In-training assessment in an undergraduate clerkship

Daelmans, H.E.M.

2005

document version
Publisher's PDF, also known as Version of record

IN-TRAINING ASSESSMENT IN AN UNDERGRADUATE CLERKSHIP

Feasibility, reliability, effect on learning environment

Hester Daelmans

ISBN: 90-9018489-4

Lay-out: Udghosh Hessel
Illustrations: Peter Korteman
Printed by: Febodruk BV

VRIJE UNIVERSITEIT

In-training assessment in an undergraduate clerkship

Feasibility, reliability, effect on learning environment

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. T. Sminia,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der Geneeskunde
op woensdag 29 juni 2005 om 13.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Hester Elisabeth Maria Daelmans

geboren te Roermond

promotoren: prof.dr. C.D.A. Stehouwer
prof.dr. C.P.M. van der Vleuten

copromotoren: prof.dr. A.J.J.A. Scherpbier
prof.dr. A.J.M. Donker

Contents

Chapter 1

Introduction

Learning in the workplace

In most undergraduate and postgraduate medical education programmes, students receive a large part of their training in the workplace. Workplace learning offers an authentic learning environment, because learning takes place in the same environment where the learners will practise their future profession. Workplace learning in undergraduate medical education predominantly occurs in clinical clerkship rotations. Two related developments in undergraduate medical education have drawn special attention to clerkship learning. Firstly, the goals of undergraduate (and postgraduate) medical education are increasingly being defined in terms of competences rather than discrete learning objectives.[1-5] Competences require integration of relevant knowledge, skills and behaviour in dealing with complex situations and problems in an appropriate manner. Because of their integrated nature, competences are best learned in an authentic learning environment.[6-8] Secondly, as the nature and sites of clinical care have changed, many medical schools have started to use community and ambulatory care settings for their clinical training programmes.[9] Thus the focus of training in undergraduate clinical clerkships today has shifted towards workplace learning as a way to facilitate competence learning and to offer students specific experiences in clinical care. As clerkship learning is becoming increasingly important in the undergraduate medical curriculum, it is also becoming increasingly important to have knowledge of how students learn in this context and of the requirements to be met by clerkships as a learning environment.

The learning environment in clerkships

A simplified model of clerkship learning shows medical students interacting with patients under the supervision of experienced clinicians. In recent years, however, the number of hospitalised patients and the length of hospital stay have decreased dramatically. In addition, patients who are admitted to hospital are often very seriously ill, which limits the possibilities for students to interact with these patients. In community care and in the out-patient clinic, there are more patients who can contribute to student learning, albeit that organisational constraints in ambulatory care may interfere with learning opportunities.[10] In the last few decades, an array of health care workers, such as members of staff, nurses, residents and other students have acted as students' supervisors. In short, the clinical setting is a highly complex learning environment that is not easy to study. It is therefore not surprising that research into clerkship learning is difficult and sparse.[9]

On clerkship rotations, the acquisition of competences generally starts with students observing others performing these competences. Gradually, students progress to 'learning by doing' and thus to performing competences themselves.[11] Key factors in the effective acquisition of competences in clerkships are adequate supervision and feedback.[12-15] The effectiveness of clerkships as a learning environment is thus highly dependent on variables like supervision and feedback.

A number of studies have indicated that supervision and feedback occur rather infrequently in undergraduate clinical training.[10,13,16,17] Moreover, supervision and feedback are often provided by professionals who are not fully qualified (residents).[17,18] Although little research has been done into the quality of feedback and its content, many authors are in agreement as to the importance of certain factors for the quality of supervision and feedback.[14] These factors are: clear agreement about objectives, good structure, and continuity. Other important factors are: reflection on performance and, for beginners, direction of performance. However, in the reality of many clinical clerkships,

supervision and feedback appear to be inadequate as regards both structure and continuity due to lack of observation, lack of scheduled time for supervision and lack of regular one-to-one contact between student and supervisor.[10,16,19] This means that an increase in the frequency, structure and continuity of supervision and feedback is a prerequisite for improving the effectiveness of clerkship as a learning environment.

Effects of assessment on the learning environment

Assessment has been described as the most powerful influence on student learning behaviours.[20,21] Moreover, the knowledge that a particular competence is likely to be assessed can lead to an increase in supervision and feedback in relation to that competence.[22] Assessment can drive learning through different mechanisms, such as content and format, the information given during an assessment and the programming of assessment in the curriculum.[23] This means that assessment is a powerful tool for manipulating the learning environment and that it can be used strategically to create the desired effects. Because assessment can affect learning through many mechanisms, the impact of assessment on the learning environment must be monitored very closely. This means that it is of vital importance to research the effects of assessment so as to establish whether it is actually steering student learning into desirable directions.[24]

Performance based assessment in clerkships

The main goal of assessment in undergraduate clinical training is to determine whether students are able to perform competences they have acquired during clerkships to a standard that justifies admittance to postgraduate training programmes. Miller described a competence pyramid conceptualising four levels of clinical competence and proposed appropriate assessment formats for each level (Figure 1).[25,26]

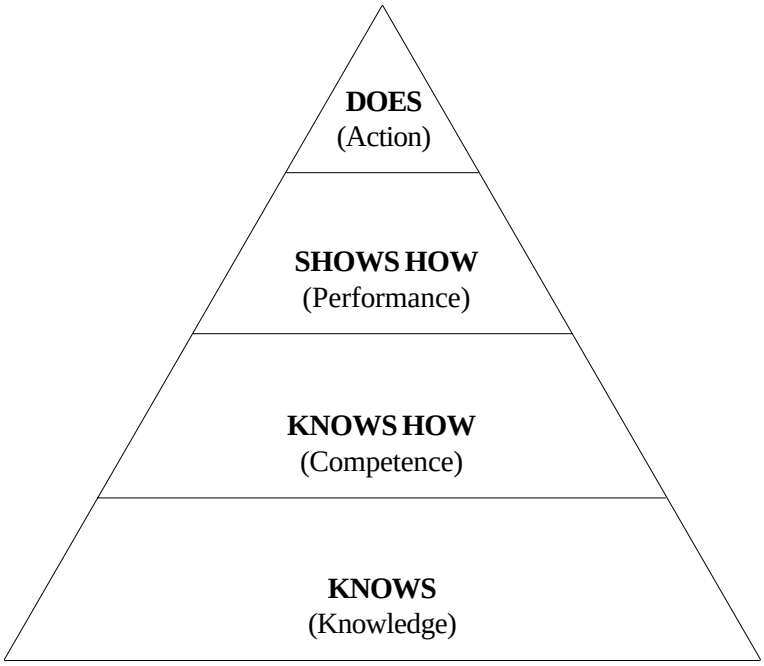


Figure 1 *Framework for clinical assessment, Miller, 1990*

The level of (isolated) theoretical knowledge ('knows'), which forms the base of the pyramid, and the next level, that of the application of knowledge ('knows how'), can be assessed by written test formats. However, these formats are not suitable for assessing actual skills performance and behaviour required for adequate performance of integrated competences. The two upper levels of Millers' pyramid are characterised as 'shows how' and 'does'. 'Shows how' refers to students demonstrating their ability to perform a competence. This level is generally assessed in simulated carefully controlled examination settings. The 'does' level of competence or the 'performance' level consists of what students or professionals actually do in authentic workplace settings. In practice, performance-based assessment at the 'does' level is difficult to implement, because it requires a careful balance between issues of reliability, validity and feasibility in real life workplace conditions.[25-28]

In the last few decades, formats used to assess student competences in clerkship settings have mainly relied on relatively unstructured global performance ratings by clinical staff as well as on long case clinical examinations.[29,30] Global performance ratings (GPRs) generally cover a number of clinically relevant competences performed by students in the course of a clerkship rotation and rated by supervisors halfway or at the end of the rotation. In the traditional long case clinical examination (long case), students are given uninterrupted and unobserved time (30-45 minutes) to take a history and perform a physical examination in a real patient, selected from the ward or outpatient clinic and who has not received any special training for the assessment. Students present their findings to the examiners as in an unstructured oral examination.[31] Generally, one or two long cases are performed at the end of a clerkship rotation. Global performance ratings and long cases have the advantage of combining feasibility with authenticity. Unfortunately, their reliability is limited due to low frequency, lack of structure, lack of direct observation of students' performance and inadequate recording of evaluators' findings. In addition, despite the high authenticity of these two assessment formats, validity remains questionable.[32,33] Finally, only a limited group of competences can be assessed using these formats.

The late seventies and eighties saw the introduction of standardised clinical examinations.[34] A standardised test format that is often used in clerkship assessment is the objective structured clinical examination (OSCE). The OSCE's strength is its good reliability, because it entails structured observation and documentation of student performance in multiple situations. However, the weakness of the OSCE lies in its feasibility, which can only be achieved at the cost of substantial investment in terms of resources and logistics.[35] Furthermore, authenticity is not one of the OSCE's strong features, given that assessments are performed in simulated contexts and often focus on isolated skills. This carries the risk of a reductionist approach to the assessment of clinical performance. It will be clear from the preceding that the 'shows how' level is the highest one at which clinical competence can be assessed by an OSCE. The OSCE is not suitable for assessing the

'does' level of performance in real practice. That is why OSCEs are generally supplemented by global performance ratings and long cases, which thus continue to make an important contribution to the assessment of clinical competence in clerkships.

Since the nineties, researchers have sought to improve the reliability and validity of traditional performance assessments (long cases and global performance ratings). Measures to do so have included: structuring assessments, increasing the numbers of both assessments and examiners, adding observation to (parts of) assessments and documenting the examiner's findings immediately following observation of a competence.[31,36-40]

For long case clinical examinations, several measures have been implemented, which resulted in better structured long cases and long cases with observation and or immediate feedback and documentation.[26,31,39] Although these measures were successful in improving the reliability and validity of long cases, the improvement in reliability was limited, with large numbers of long cases being needed to reach acceptable reliability. This meant that increased reliability could only be achieved at the expense of reduced feasibility. Long case clinical examination formats that combine good reliability and good feasibility were needed to make this assessment method viable.

The need to implement measures to improve the quality of formats of global performance rating engendered a new approach to this type of assessment under the name of in-training assessment. In-training assessment is defined as multiple, structured and observed assessments of student performance in the workplace, documented immediately after observation and spread throughout a clerkship rotation.[40-42] In-training assessment thus entails multiple observed assessments, which enable structured feedback. Typically, in-training assessment can be integrated into the day-to-day activities of students and examiners in both undergraduate and postgraduate training programmes. However, for in-training assessment to be effective, more research is needed to develop formats with good reliability and acceptable feasibility, which will enable broad implementation of this assessment method.

General themes in this thesis

In order to improve the effectiveness of undergraduate clerkships as a learning environment for the acquisition of clinical competences, changes are needed in the frequency and quality of both supervision of and feedback on students' performance of competences. In order to improve the quality of performance based assessment in undergraduate clerkships, we need to develop feasible, reliable and valid methods for the assessment of clinical competence at the 'shows how' and 'does' levels of Miller's pyramid.[25] From the perspective that assessment exerts a potentially strong steering effect on learning, this thesis explores whether the introduction of an in-training assessment programme in an undergraduate clerkship enhances the learning environment of the clerkship. Furthermore, this thesis explores reliability and feasibility aspects of the in-training assessment programme.

Research questions

This thesis explores whether introducing an in-training assessment programme into an undergraduate clerkship rotation has a positive effect on the learning environment of the clerkship. The in-training assessment programme studied was introduced to achieve feasible and reliable performance based assessment for a broad range of competences. Because the assessment programme needed to be established before its effects on the learning environment could be studied, issues regarding the feasibility and reliability of this in-training assessment programme in an undergraduate clerkship are addressed first. The context in which the studies were performed and the details of the assessment formats are described in the appendix.

Specific research questions

1. What is the feasibility and reliability of multiple long case clinical examinations in an undergraduate clinical clerkship? (chapter 2)

2. What is the feasibility and reliability of the in-training assessment programme for assessing students' clinical performance? (chapter 3)
3. What is the reliability and validity of global performance ratings in an undergraduate clerkship with increased rater-student interactions? (chapter 4)
4. What are the frequencies of supervision, feedback and assessment regarding a set of specified competences and what are the differences between disciplines? (chapter 5)
5. What is the effect of the in-training assessment programme on the frequency of supervision and the quality of feedback on clinical competences and on inter-student differences in these respects? (chapter 6)
6. How is the in-training assessment programme actually carried out (curriculum in action) and what are the main features of supervision and feedback in the undergraduate clinical clerkship in which the in-training assessment programme is integrated? (chapter 7)

Chapters two through seven present studies that we conducted to answer the different research questions. As each chapter is based on a journal article, some repetition was inevitable.

Chapter two addresses the feasibility and reliability of a long case clinical examination format comprising multiple structured examinations and multiple examiners. This assessment format was named 'multiple oral examinations', because students' competence was assessed when they presented their findings to the examiners as in an oral examination. However, the student patient encounter preceding the oral examination was the crucial component of this assessment format. Therefore, in the following chapters, this assessment format will be called multiple long case clinical examination.

Chapter three presents a study of the feasibility and reliability of the in-training assessment programme, a performance based assessment programme comprising five test formats. The multiple long case clinical examination format described in chapter two was included in the in-training assessment programme as a multiple sample test.

Chapter four deals with multiple global performance ratings (GPRs) in an undergraduate setting with extensive rater-student interactions. In addition to reliability, the predictive validity of these global performance ratings for the in-training assessment programme (criterion variable) was studied.

Chapter five describes the frequency of supervision, feedback and assessment for a set of specified competences across three different undergraduate clinical clerkships in which traditional assessment formats were used (one long case clinical examination and one global performance rating).

Chapter six describes how the implementation of the in-training assessment programme in one of the undergraduate clinical clerkships affected the frequency and quality of supervision and feedback.

Chapter seven presents a qualitative study of the in-training assessment programme in action and its qualitative effects on supervision and feedback.

Chapter eight synthesises the findings, considers the methodology that was used and discusses the answers to the different research questions. It also presents some recommendations for further research.

Finally, this thesis is summarised in English and in Dutch.

References

1. The Medical School Objectives Writing Group. Learning objectives for medical students education - guidelines for medical schools: Report I of the Medical School Objectives Project. *Acad Med* 1999;74:13-8.
2. CANMEDS. Skills for the new millennium. CANMEDS 2000 project. http:/www.rcpsc.medical.org/publications/index.php. 2000. Accessed Januari 2005.
3. WFME Task Force on Defining International Standards in Basic Medical Education. Report of the working party, Copenhagen, 14-16 October 1999. *Med Educ* 2000;34:665-75.
4. Metz JCM, Stoelinga GBA, Pels Rijcken-van Erp Taalman Kip EH, van den Brand-Valkenburg BWM. *Blueprint 1994: Training of doctors in the Netherlands*. Objectives of undergraduate medical education in the Netherlands. Nijmegen: University Publication Office 1994.
5. Karle H. Global standards in medical education - an instrument in quality improvement. *Med Educ* 2002;36:604-5.
6. American Health Information Management Association. Position statement. Issue: Educational clinical affiliations. *Journal of AHIMA* 1995;66: suppl 1-2 following p. 104.
7. Headrick LA, Neuhauser D, Schwab P, Stevens DP. Continuous quality improvement and the education of the generalist physician. *Acad Med* 1995;70:S104-9.
8. Regehr G, Norman GR. Issues in cognitive psychology: implications for professional education. *Acad Med* 1996;71:988-1001.
9. Woolliscroft JO. Medical student clinical education. In: Norman GR, van der Vleuten CPM, Newble DI, eds. *International Handbook of Research in Medical Education.* Dordrecht/Boston/London: Kluwer Academic Publishers 2002;365-80.
10. Irby DM. Teaching and learning in ambulatory care settings: a thematic review of the literature. *Acad Med* 1995;70:898-931.
11. Lave J, Wenger E. *Situated Learning: Legitimate Peripheral Participation.* Cambridge,UK: Cambridge University Press 1991.
12. Irby DM. What clinical teachers in medicine need to know. *Acad Med* 1994;69:333-42.
13. Jolly BC, Macdonald MM. Education for practice: the role of practical experience in undergraduate and general clinical training. *Med Educ* 1989;23:189-95.
14. Kilminster S, Jolly B, van der Vleuten CPM. A framework for effective training for supervisors. *Med Teach* 2002;24:385-9.
15. Rolfe I, Sanson-Fisher RW. Translating learning principles into practice: a new strategy for learning skills. *Med Educ* 2002;36:345-52.

16. Remmen R, Denekens J, Scherpbier A, Hermann I, van der Vleuten CPM, van Royen P, Bossaert L. An evaluation study on the didactic quality of clerkships. *Med Educ* 2000;34:460-4.
17. Van der Hem-Stokroos HH, Scherpbier AJJA, van der Vleuten CPM, de Vries H, Haarman HJ. How effective is a clerkship as a learning environment? *Med Teach* 2001;23:599-604.
18. Remmen R, Denekens J, Scherpbier AJJA, van der Vleuten CPM, Hermann I, Puymbroeck van H, Bossaert L. Evaluation of skills training during clerkships using student focus groups. *Med Teach* 1998;20:428-31.
19. Branch WT, Paranjape A. Feedback and reflection: teaching methods for clinical settings. *Acad Med* 2002;77:1185-8.
20. Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. *Med Educ* 1983;17:165-71.
21. Messick S. The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* 1994;23:13-23.
22. Stillman PL, Haley HL, Regan MB, Philbin MM. Positive effects of a clinical performance assessment program. *Acad Med* 1991;66:481-3.
23. Van der Vleuten CPM. The assessment of professional competence: Development, research and practical implications. *Adv Health Sci Educ Theory Pract* 1996;1:41-67.
24. Van Luijk SJ, van der Vleuten CPM, Schelven RM. The relation between content and psychometric characteristics in performance-based testing. In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP, eds. *Teaching and Assessing Clinical Competence.* Groningen: Boekwerk Publications 1990;202-7.
25. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990;65:S63-7.
26. Wass V, van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. *The Lancet* 2001;357:945-9.
27. Van der Vleuten CPM, Scherpbier AJJA, Dolmans DHJM, Schuwirth LWT, Verwijnen GM, Wolfhagen HAP. Clerkship assessment assessed. *Med Teach* 2000;22:592-600.
28. Schuwirth LWT, van der Vleuten CPM. Changing education, changing assessment, changing research? *Med Educ* 2004;38:805-12.
29. Magarian GJ, Mazur DJ. Evaluation of students in medicine clerkships. *Acad Med* 1990;65:341-5.
30. Kassebaum DG, Eaglen RH. Shortcomings in the evaluation of students' clinical skills and behavior in medical school. *Acad Med* 1999;74:842-9.
31. Wass V, Jolly B. Does observation add to the validity of the long case? *Med Educ* 2001;35:729-34.
32. Van der Vleuten CPM. Validity of final examinations in undergraduate medical training. *BMJ* 2000;321:1217-9.
33. Wass V, van der Vleuten CPM. The long case. *Med Educ* 2004;38:1176-80.

34. Harden RM, Gleeson FA. ASME Medical Education Booklet no. 8. Assessment of medical competence using objective structured clinical examination (OSCE). *J Med Educ* 1979;13:41-54.
35. Mavis BE, Henry RC, Ogle KS, Hoppe RB. The emperor's new clothes: the OSCE reassessed. *Acad Med* 1996;71:447-53.
36. Anastakis DJ, Cohen R, Reznick RK. The structured oral examination as a method for assessing surgical residents. *Am J Surg* 1991;162:67-70.
37. Elks ML, Sawyer JR. A format for scripted clinical oral examinations. *Med Educ* 1993;27:160-4.
38. Schwiebert P, Davis A. Increasing inter-rater agreement on a family medicine clerkship oral examination-a pilot study. *Fam Med* 1993;25:182-5.
39. Gleeson F. The effect of immediate feedback on clinical skills using the OSLER. In: Rothman AI, Cohen R, eds. *Proceedings of the Sixth Ottawa Conference on Medical Education.* Toronto: University of Toronto Bookstore Custom Publishing 1994;412-5.
40. Turnbull J, MacFadyen J, van Barneveld C, Norman G. Clinical work sampling: a new approach to the problem of in training evaluation. *J Gen Int Med* 2000;15:556-61.
41. Feletti G, Cameron D, Dawson-Saunders B, des Groseilliers JP, Dooley B, Farmer E, McAvoy P. In-training assessment. In: Newble D, Jolly B, Wakeford R, eds. *The Certification and Recertification of Doctors: Issues in the Assessment of Clinical Competence.* Cambridge: Cambridge University Press 1994;151-66.
42. Turnbull J, van Barneveld C. Assessment of clinical performance: in-training evaluation. In: Norman GR, van der Vleuten CPM, Newble DI, eds. *International Handbook of Research in Medical Education.* Dordrecht/ Boston/London: Kluwer Academic Publishers 2002;793-810.

Appendix

In-training assessment programme in this thesis

Traditional clerkships in the Vrije Universiteit Medical Center

At theVrije Universiteit Medical Center (VUmc), Amsterdam, the Netherlands, assessment in undergraduate clerkships traditionally consisted of two assessment formats, i.e. a global performance rating and a long case clinical examination. In most clerkships, assessment entailed one end-of-clerkship global performance rating and one long case clinical examination in the last week of the rotation or after the rotation was finished. Because the traditional performance based assessment method was far from optimal, it was decided to reorganise clerkship assessment.

At the VUmc, students enter clerkships after four preclinical years of interdisciplinary, mainly lecture-based education and a six-week course in the skills training laboratory. The first discipline-based clerkship was the Internal Medicine clerkship. Although the course in the skills training laboratory offered specific preparation for entering the clerkship, students were found to experience great difficulties in the first few weeks of the Internal Medicine clerkship, such as in performing a history and physical examination in an old and very sick patient on the ward within a limited amount of time and also in dealing with medical and other health care staff. Students needed a lot of supervision to overcome these difficulties. The fact that supervision could not always be provided meant that in many cases less than optimal use was made of the learning opportunities offered in the first few weeks of the Internal Medicine rotation.

Reorganisation of the clerkships in the Vrije Universiteit Medical Center

The reorganisation of clerkship assessment comprised the following elements. Firstly, in the Internal Medicine clerkship at the VUmc, the single long case clinical examination was replaced by a multiple long case clinical examination. In addition, measures were implemented to improve the reliability of the long case clinical examination, for instance by increasing the number of examiners and by structuring the rating procedure. Secondly, an in-training assessment programme, comprising a student-patient encounter, a critical appraisal session, a case presentation, case write-ups and multiple long case clinical examinations was introduced in all VUmc clerkships. Thirdly, structured rating forms were used to rate students' performance and record verbal feedback. Fourthly, the structure of the global performance rating was changed in such a way that the ratings of the in-training assessment programme could be included in the global performance ratings.

In order to provide the extensive supervision that students obviously needed during the first weeks of their Internal Medicine clerkship, a new, three-week clerkship (the introductory clerkship) with frequent and regular supervision was introduced. This clerkship was scheduled before the Internal Medicine clerkship.

Elements of the reorganisation that were studied in this thesis

The following aspects relating to the reorganisation of clerkship assessment are addressed in this thesis: 1) the feasibility and reliability of the multiple long case clinical examination format (chapter 2); 2) the feasibility and reliability of the in training assessment programme (chapter 3); and 3) the reliability and validity of the global performance rating in the introductory clerkship (chapter 4). Before describing these three assessment formats, we present background information in which we describe the considerations underlying the construction of the formats.

Long case clinical examinations

The measures we took to ensure the feasibility of the long case were: implementing long cases in day- to-day work in the outpatient clinic, spreading the long cases over the students' final week in the outpatient clinic (two per day for a maximum of five consecutive days) and appointing one examiner for one day in that week who was the member of staff in charge for the student on that day. The measures put in place to improve the reliability of the long case clinical examination were: increasing the number of examinations, increasing the number of examiners and structuring the assessment by means of a structured rating form.

The long case clinical examination consisted of a patient interview and physical examination performed by the student, who subsequently had to write a medical record in a fixed amount of time. The clinical examiner verified the findings of the interview and the clinical examination and could ask the student to demonstrate specific clinical skills. After that, the clinical examiner studied the medical record and asked mainly case-related questions. The examiner rated student performance on a list of items concerning the different stages of the clinical encounter and the quality of the discussion. In addition, a global rating was given for the student's overall clinical performance.

In-training assessment programme

The in-training assessment programme that was studied in this thesis contained multiple test formats. This design was based on the consideration that a variety of methods was needed, chosen on the basis of their effectiveness in assessing particular competences.[1] Five test formats, together covering a broad range of competences, were incorporated into the programme. In designing the in-training assessment programme, special consideration was given to the feasibility of every format. The required frequency of the selected formats was determined by compromising between sampling needs to overcome domain specificity on the one hand and logistics and

resource requirements (feasibility) on the other hand. Multiple structured long cases with systematic feedback and careful documentation, which were embedded in the clerkship were regarded as in-training assessment and thus incorporated into the in-training assessment programme. The in-training assessment programme was implemented specifically for the purpose of improving clerkship learning. It was assumed that it would lead to increased frequency, structure and continuity of supervision and feedback on the assessed competences and thus to improvement of the learning environment of the clerkship.

The in-training assessment programme comprised one student-patient encounter, one critical appraisal session, one case presentation, four structured long cases and twelve case write-ups. Members of staff acted as examiners for the first four formats and residents assessed the case write-ups. Structured rating forms with a five-point Likert scale were used to rate student performance and record verbal feedback.

Global performance rating

The following measures were taken to enhance the feasibility of global performance ratings: using a single-item rating form, providing examiners with a computerised form and a scanned picture of the examinee and enabling examiners to complete and return the form by computer. Measures implemented to improve the reliability of global performance ratings were: structuring the clerkship in such a way that students were regularly supervised by examiners and increasing the number of examiners who performed the global performance ratings.

The global performance ratings were given at the end of the introductory clerkship and comprised a single item rating on a digital form.

Reference

1. Schuwirth LW, Southgate L, Page GG, Paget NS, Lescop JM, Lew SR, Wade WB, Baron-Maldonado M. When enough is enough: a conceptual basis for fair and defensible practice performance assessment. *Med Educ* 2002; 36:925-30.

Chapter 2

Reliability of clinical oral examinations re-examined

HEM Daelmans, AJJA Scherpbier, CPM van der Vleuten, AJM Donker.

Published in *Medical Teacher, 2001;23:422-424*

Summary

Many medical schools still use oral examinations for the evaluation of clinical competence of students in their clerkship, although it has been proven that orals have poor reliability. This study investigates the feasibility and reliability of multiple oral examinations. Students in the last week of their Internal Medicine clerkship in an outpatient clinic were given several patient-based oral examinations. The student's performance was rated on a list of items reflecting clinical competence. A global judgement of the student's performance was also given. The results indicate that it is possible to increase the number of orals and the number of examiners in the day-to-day practice of an outpatient clinic moderately. The reliability when using a number of orals is better than the reliability of the common single oral examination. The reliability using global judgements appeared to be better than the reliability of averaged item scores.

Overview

The reliability of oral examinations need not be poor or at least no less poor than other clinical assessment methods.

Moderately increasing the number of oral examinations and number of examiners improves the reliability substantially and is feasible in the daily practice of an outpatient clinic.

The reliability of global judgements appeared to be better than the reliability of averaged item scores.

Introduction

Oral examinations for the evaluation of clinical competence of students in their clinical rotations have been shown to have poor reliability.[1,2] Although traditional oral examinations are still used by many medical schools as a method of student assessment, other ways of assessing clinical competence have been explored. More structured orals and longer testing time per case have been proposed to improve the reliability.[3-5] In stead of orals, other clinical evaluation instruments such as the objective structured clinical examination (OSCE) or the simulated patient (SP)-based test have been developed.[6,7] The reliability of these instruments is, however, equally problematic, mainly due to the problem of content specificity of clinical competence. These performance based tests require large samples of clinical performance from each candidate and relatively long testing time before reliable information is obtained.[8] Oral examinations are usually short and only sample a small content area (e.g., one patient case). We also know that the reliability of oral examinations is seriously threatened by examiner variation. Orals typically have a small sample of examiners (e.g., one or two) and lack in inter-rater reliability.[1,8] Very often oral examinations are graded globally using no or very little items to structure the grading. The poor reliability of orals could therefore be explained by the small samples of examiners, the small sample of content or cases, and the global judgements that are used. Perhaps the reliability of orals compared with other clinical competence measures is equally good (or poor) when larger samples are used.

The purpose of this study was first, to investigate the feasibility of an oral examination when numbers of orals and examiners are increased and second, to estimate the reliability of the oral examinations as a function of the number of orals. The scoring of candidate performance was carried out by rating the performance on a number of items and by a single global judgement. The reliability of the averaged item scores and the global judgement score was compared. During the last week in the

outpatient clinic at the internal medicine rotation, multiple oral examinations were conducted by different (single) examiners every day.

Method and material

Context of the study

The study was conducted at the Department of Internal Medicine of the Vrije Universiteit in Amsterdam, The Netherlands. The medical curriculum at the Vrije Universiteit consists of four preclinical years and two clinical years during which the students rotate in clerkships. In the first ten weeks, during the rotation at the Department of Internal Medicine, students stay at the in-patient clinic and are assigned to residents for the day-to-day clinical work. The last four weeks are spent at the outpatient clinic where the students can interview and examine at least two patients each day. During the entire rotation the students are supervised by members of staff.

Procedure

A patient-based oral examination twice a day for a maximum of five days in a row with a (single) different examiner each day was planned for students in the last week of their Internal Medicine clerkship in the outpatient clinic. The examination was incorporated in the day-to-day work of both student and examiner, and conducted by the member of staff in charge for the student that day. Five examiners were involved in the study. Each examiner was responsible for two orals per student one day a week. The interaction between examiner and student during one oral examination lasted approximately 30 minutes. The time for the student-patient contact (1,5 hours) and the doctor-patient contact (approximately 15 minutes) when followed by an oral examination equalled the time of these contacts when not followed by an oral examination.

The examination consisted of an interview and a physical examination of a patient by the student, who was subsequently

required to write a medical record in a fixed amount of time. The medical record was submitted to the examiner, who verified the findings of the interview and the clinical examination. The examiner could also ask the student to demonstrate specific clinical skills. Although the examiner could observe the student during the patient contact, this hardly occurred. After completing the patient contact, the examiner interviewed the student asking case-related questions concerning, for example, findings of the interview and the physical examination, differential diagnosis and therapy. The examiner could also ask questions to test general knowledge. The examiner rated the performance on a list of items referring to the stages of the clinical contact and the quality of the discussion. The rating was based mainly on performance during the interview of the student by the examiner, the actual oral examination. The items on which a grade could be given were: history taking, physical examination, medical record, summary, differential diagnosis, additional investigations, therapy and discussion of the case. For every item a grade was given on a 10 point scale (with a score of 5 or less being unsatisfactory). In addition, a global judgement for the integral clinical performance of the student was given, again on a 10-point scale.

Statistical analysis

The feasibility of raising the number of orals and examiners per students was studied by calculating the frequency of orals and examiners per student. To estimate the reliability of the oral examinations as a function of the numbers of orals, a generalisability study using a simple person X oral design was used. Since the context varied from oral to oral, and because five examiners graded the orals, orals were considered nested within persons (o:p design). After variance component estimation, generalisability coefficients were estimated as a function of the actual and projected number of orals. These analyses were carried out separately on the averaged item score, the global judgement, and the averaged combination of these scores.

Table 1
Frequency of realised orals

Number of orals	1	2	3	4	5	6	7	8	9	10
Number of students	-	3	1	3	6	10	8	9	4	8

Results

Fifty-two students were included in the study. Table 1 shows the frequency of orals per student. Although students were required to take two examinations for five days with a different (single) examiner each day, most of the students were examined on fewer days and by fewer examiners. On some days, students were examined only once. Thirty-nine students (75%) took at least six orals. The database of these students was used for further analysis.

Table 2
Descriptive statistics on the grades that were given in the group of 39 students who received at least six examinations

	Mean	**SD**
Averaged items	6.60	0.58
Global judgement	7.27	0.52
Averaged sum of both	6.94	0.52

Table 2 presents the mean frequencies of the grades in this group of 39 students. The grades on the global judgement were significantly higher ($p<0.001$) than the averaged item grades. The correlation between the averaged sum of both scores (global and averaged items) and the global grades is significant, but not very high (0.59 $p=0.01$).

This indicates that the ranking of students is different to some extent for both grades.

Table 3 shows the reliability of the oral examinations as a function of the numbers of orals for averaged item grades, the global judgement, and the mean of both previous scores. The generalisability coefficient for six examinations is 0.62 for the averaged item grade and 0.72 for the global judgement score and the mean of both scores. These generalisability coefficients reflect both inter-examiner variation and inter-case variation. Owing to the limitation of the design (one examiner per oral), these sources of variation cannot be disentangled. A reliability of 0.8 can be reached with about 15 orals (7,5 hours) for the averaged item grades and with nine orals (4,5 hours) for the global judgement or the mean of both scores.

Table 3
Generalisability coefficients of the oral examination as a function of the number of patient cases based on the averaged item grades and on the global judgement

Number of cases (Testing time in hours)	**1** (0.5)	**2** (1.0)	**3** (1.5)	**4** (2.0)	**6** (3.0)	**8** (4.0)	**10** (5.0)	**12** (6.0)	**14** (7.0)	**16** (8.0)
Averaged item grades	.22	.35	.45	.52	.62	.69	.73	.77	.79	.81
Global judgement	.30	.47	.57	.64	.72	.78	.81	.84	.86	.87
Averaged sum of both	.30	.46	.56	.63	.72	.77	.81	.84	.86	.87

Discussion

It was possible to moderately increase the number of orals and the number of examiners in the day-to-day practice of an outpatient clinic. Two orals for five days in a row by a single different examiner every day was apparently too ambitious a goal in the day-to-day practice of an outpatient clinic. Six to eight orals conducted by four to five examiners during one week was feasible in practise.

Raising the number of orals and the number of examiners (and thereby increasing the examination time, the number of examiners and the range of patient problems) increased the reliability of the examination, compared to one oral conducted by one examiner. The single oral examination, common in many clerkships, is virtually hopeless in terms of reliability.

The reliability of the oral examination with a small number of cases and examiners remained poor. At least five to eight hours of testing time was required before an adequate reliability was achieved. However, this is comparable to the OSCE and to another study focused on oral examinations.[8,9] The OSCE, though, is an examination that cannot be incorporated in a day-to-day routine of an outpatient clinic, whereas the multiple oral examination can take place in such a setting without modifications.

There is a difference between the reliability of the averaged item grade and the global judgement. The global judgement reliability was better than the averaged item grade reliability. This indicates that the global judgement uses additional information on top of that covered by the items. Apparently some other aspects of the student's functioning are also taken into account. These aspects presumably cannot be assessed using the individual items. Van Luijk *et al.* and Cunnington *et al.* found that reliabilities of global rating scales were as good as or better than those of checklists.[10,11] Littlefield and Troendle found that combination of global judgement and detailed assessments improved reliability.[12] In our study, combining global judgement and averaged item grades did not further improve reliability.

The limitations of this study are at least twofold. First, a small samples size of students is used. For stable estimates of variance components a larger sample would be preferred. The findings of this study need extension through further research. This is particularly desirable because most reliability studies reported so far have focussed exclusively on inter-rater reliability, thereby ignoring inter-case variability which is actually a dominant error source in clinical competence assessment. Second, the design assumed full nesting of examiners within orals (i.e. a different examiner for each oral).

However, most examiners conducted two orals per student. This may have caused some overestimation in the generalisability coefficient.

Conclusion

A moderate increase in the number of oral examinations and number of examiners is feasible in the day-to-day practice of an outpatient clinic. The reliability of using a number of orals is substantially better than the commonly used single oral examination. The reliability of global holistic assessments is better than the reliability of more analytic assessments. However, many orals are needed in order to achieve a reliable score, requiring at least nine to ten occasions or five hours of testing time. Although multiple orals can be combined with daily work in an outpatient clinic, in most medical schools it is probably impossible to achieve nine to ten occasions on a structural basis without taking additional measures. One should realise, however, that this is by no means, different than any other measure of clinical competence. For that matter, the reliability of the oral examination is not poor, or at least not less poor than any other clinical assessment method.

References

1. Weingarten MA, Polliack MR, Tabenkin H, Kahan E. Variations among examiners in family medicine residency board oral examinations. *Med Educ* 2000;34:13-7.
2. Muzzin LJ, Hart L. Oral examinations. In: Neufeld VR, Norman GR, eds. *Assessing Clinical Competence.* New York: Springer Publishing Co 1985; 71-93.
3. Anastakis DJ, Cohen R, Reznick RK. The structured oral examination as a method for assessing surgical residents. *Am J Surg* 1991;162:67-70.
4. Schwiebert P, Davis A. Increasing inter-rater agreement on a family medicine clerkship oral examination-a pilot study. *Fam Med* 1993;25:182-5.
5. Elks ML, Sawyer JR. A format for scripted clinical oral examinations. *Med Educ* 1993;27:160-4.
6. Hart IR, Harden RM, Walton H, eds. *Newer Developments in Assessing Clinical Competence*. Montreal: Can Health Publications 1986.
7. Hart IR, Harden RM, eds. *Further Developments in Assessing Clinical Competence*. Montreal: Can Health Publications 1987.
8. Van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardised patients: State of the art. *Teach Learn Med* 1990;2:58-76.
9. Swanson DB. A measurement framework for performance-based tests. In: Hart IR, Harden RM, eds. *Further Developments in Assessing Clinical Competence*. Montreal: Can Health Publications 1987;13-45.
10. Van Luijk SJ, Van der Vleuten CPM. A comparison of checklists and rating scales in performance-based testing. In: Hart IR, Harden RM, DesMarchais J, eds. *Current Developments in Assessing Clinical Competence.* Montreal: Can Health Publications 1992;357-82.
11. Cunnington JPW, Neville AJ, Norman GR. The risk of thoroughness: reliability and validity of global ratings and checklists in an OSCE. *Adv Health Sci Educ*1997;1:227-33.
12. Littlefield J, Troendle R. Global judgement in evaluating dental student clinical performance. *Dent Sch Q* 1987;3:11-4.

Chapter 3

Feasibility and reliability of an in-training assessment programme in an undergraduate clerkship

HEM Daelmans, HH van der Hem-Stokroos, RJI Hoogenboom, AJJA Scherpbier, CDA Stehouwer, CPM van der Vleuten.

Published in *Medical Education 2004;38:1270-1277*

Summary

Structured assessment, embedded in a training programme, with systematic observation, feedback and appropriate documentation may improve the reliability of clinical assessment. This type of assessment formats is referred to as in-training assessment (ITA). The feasibility and reliability of an ITA programme in an internal medicine clerkship were evaluated. The programme comprised four ward-based test formats and one outpatient clinic-based test format. Of the four ward-based test formats three were single-sample tests, consisting of one student-patient encounter, one critical appraisal session and one case presentation. The other ward-based test and the outpatient-based test were multiple sample tests, consisting of 12 ward-based case write-ups and four long cases in the outpatient clinic. In all the ITA programme consisted of 19 assessments. During 41 months, data were collected from 119 clerks. Feasibility was defined as over two thirds of the students obtaining 19 assessments. Reliability was estimated by performing generalisability analyses with 19 assessments as items and with five test formats as items.
A total of seventy-three student (69%) completed 19 assessments. Reliability expressed by the generalisability coefficients was 0.81 for 19 assessments and 0.55 for five test formats.
The ITA programme proved to be feasible. Feasibility may be improved by scheduling protected time for assessment for both students and staff. Reliability may be improved by more frequent use of some of the test formats.

Overview

What is already known

In-training assessment (ITA) is beneficial to the reliability of clinical performance evaluation
The feasibility of ITA may be limited.

What this study adds

An ITA programme comprising a variety of test formats for assessing undergraduate students' performance is feasible.
An ITA programme can reach acceptable reliability by careful sampling across and within test formats.

Suggestions for further research

Challenges for research into ITA is to find the right balance between methods and ways to combine information across methods

Introduction

During the last few decades various methods have been devised to evaluate the clinical competence of undergraduate medical students.[1] Traditionally, knowledge has been the focal point of clinical assessment. Although knowledge is undeniably a vital component of clinical competence and thus must be assessed, true clinical competence depends on the ability to integrate knowledge, skills and attitudes. This means that students' integrative capacity should be a prominent feature of clerkship assessment.[2] The formats used to assess students' performance in clinical practice have mainly relied on relatively unstructured evaluation by clinical staff as well as on long cases.[3,4] These formats have the advantage of combining feasibility with authenticity, but their reliability is low as a result of low frequency and lack of direct observation of students' performance. Additional weaknesses of clinical

evaluation are lack of structure of the evaluation programme and inadequate recording of evaluators' findings. Moreover, in the assessment no formal distinction is made between different clinical skills, such as diagnostic reasoning skills and communication skills.[5]

The late 1970s and the 1980s saw efforts to resolve the drawbacks of the traditional assessment formats by the introduction of standardised clinical examinations such as the objective structured clinical examination (OSCE).[6] These formats replaced or supplemented the traditional tests. The strength of the OSCE is structured observation and documentation of students' performance in multiple situations. Feasibility is problematic, however, due to the huge investment in terms of the resources and logistics needed for reliable assessment.[7] Authenticity is also not one of the OSCE's strong points, with stations predominantly testing skills in isolation and in simulated contexts. The inevitable conclusion was that innovative assessment formats would have to be devised in order to achieve reliable assessment of integrated clinical knowledge, skills and attitudes in an authentic context.[5]

During the 1990s attention was focused on the reliability of clerkship assessment in an authentic context. First, the reliability of assessment of students' performance during clerkships and the ability of assessment to distinguish between competences was improved by the implementation of more structured assessment, documentation of behaviour immediately following observation and increased numbers of assessments and examiners.[8,9] The term 'in-training assessment' (ITA) is used to refer to systematic observation with feedback and documentation of students' performance integrated in a training programme (both undergraduate and postgraduate training).[9,10] The reliability of the long case and its ability to comprehensively evaluate competences were enhanced through better structure and more cases and/or examiners.[11-13] Multiple, structured long cases, embedded in a clinical attachment, with systematic feedback and careful documentation can be characterised as ITA. The ITA approach to clinical assessment appears to have a great deal of potential for

improving reliability and distinguishing between different competences. However, the effectiveness of ITA depends crucially on direct observation of students' performance followed by careful recording of the findings. These requirements pose a feasibility problem, which may form a barrier to the realisation of the full potential of ITA.[9,13] Nevertheless, ITA appears to be a promising approach to clinical assessment. This is why we designed and piloted an ITA programme in the general internal medicine clerkship at the Vrije Universiteit Medical Centre, Amsterdam, the Netherlands.

The ITA programme was comprised of four assessment formats to be used for ward-based assessment (student-patient encounter, critical appraisal session, case presentation and case write-up) and one format (structured long case) to be used in the outpatient clinic. The programme was implemented in the internal medicine clerkship. Efforts were made to maximise the feasibility of the programme by careful planning of the frequency of assessment and by planning assessment to fit into students' and examiners' schedules. Observation or discussion immediately followed by the recording of the assessment results was planned for all the components of the ITA programme. The study reported in this paper addresses the feasibility and reliability of the ITA programme for assessing the clinical performance of students during the internal medicine clerkship.

Method

In-training assessment programme

Blueprinting

First we determined which clinical competences were to be included in the ITA programme. The first and second author selected the most relevant competences from the Dutch national blueprint of undergraduate medical education: history taking and physical examination skills, diagnostic reasoning skills,

management skills, verbal and written communication skills, critical appraisal skills and professional skills.[14] This selection is comparable with competence or outcome lists published in other countries.[15-17]

Subsequently, test formats were identified in consideration of our requirement for a variety of methods, each chosen on the basis of its effectiveness in assessing a particular competency.[18,19] The selected formats are described below. Finally, we decided on the required frequency of the selected formats. This was the benchmark for the feasibility of the ITA programme. The estimated frequencies were arrived at by compromising between the sampling needs to overcome the domain specificity of many competences on the one hand, and considerations regarding logistics and resource requirements on the other.

In-training test formats

Student-patient encounter

A staff member observed a student patient encounter, assessed history taking and physical examination skills, diagnostic reasoning skills and verbal communication skills, and recorded the findings on a rating scale. The reason for including this assessment format was that students generally receive very little direct feedback when they interview and examine patients.[20] What feedback students receive is mostly provided in brief contacts between students and staff which generally do not coincide with the student-patient encounter.[21] We had to compromise with regard to sampling needs, given the considerable demands of this type of assessment on staff in terms of time and planning. We therefore required one student-patient encounter.

Critical appraisal session

A staff member observed students presenting a review of a scientific paper on a clinical topic to a group of students. The scientific relevance of the content of the presentation and the quality of critical thinking, verbal communication and presentation skills were rated. Because these sessions require the attendance of a staff member as well as the entire student group, one presentation per student was considered to be the maximum number feasible. Moreover, students' communication and presentation skills were also evaluated in case presentations.

Case presentation

A staff member observed a student presenting a patient case to a group of students and staff. The completeness and relevance of the information presented was rated as well as the quality of critical thinking, verbal communication and presentation skills. The limitation to one presentation per student was based on the same time constraints and similarity considerations that influenced the restriction of the number of critical appraisals.

Case write-up

Clerks are mostly supervised by doctors who are receiving their postgraduate specialised training (residents). The resident who supervised the student during day-to-day clinical work discussed the student's case write-up with the student and evaluated diagnostic reasoning skills, management skills and written communication skills. Discussing students' diagnostic reasoning and management skills is among the regular daily interactions between resident and undergraduate student. We designed the case write-ups to enable students to reconsider their diagnostic reasoning and management plan for a specific patient and practise writing a summary of the main findings. This test format served the additional purpose of helping residents structure the discussion and

rating of students' diagnostic reasoning skills, management skills and written communication skills. At least 12 case write-ups had to be evaluated for each student. This high frequency is justified because written communication skills are notoriously poorly supervised, diagnostic reasoning and management skills are highly domain-specific and the format can easily be fitted into the daily routine of residents' supervisory tasks.[22 23]

Structured long case

A student and a staff member met for 30 minutes to discuss the student's write-up of a complete interview and physical examination in the outpatient clinic. Items to be rated were history taking and physical examination skills, diagnostic reasoning skills, management skills, written communication skills and critical appraisal skills. We included the structured long case in the ITA programme because it is still widely used by many examiners and institutions, who value it as an important tool to evaluate students' clinical performance.[24,25] Although observation improves the validity of long cases as an assessment instrument, we decided to use unobserved cases because mandatory observation was expected to be impracticable in a busy outpatient clinic.[26] Four structured long cases were required for every student. Earlier studies had shown this to be a feasible number in the outpatient clinic, although the reliability of four long cases as a single assessment method was found to be inadequate.[1,13]

In summary, the ITA programme for the internal medicine clerkship consisted of five test formats. A minimum frequency was specified for each format, yielding a required total of 19 assessments comprising three single-sample formats (student-patient encounter, critical appraisal session and case presentation) and two multiple-sample formats (12 case write-ups and four structured long cases). Each test format was administered and evaluated by a different group of examiners. Multiple sample formats were assessed by at least two examiners.

Rating forms

On their first day in the internal medicine department students were provided with a set of rating forms for the different test formats to familiarise them with the assessment programme. Students were instructed to give the appropriate form to the examiners at the time of the assessment. The student was to keep the completed forms until the end of the clerkship.

All forms used a five-point Likert scale (1 = fail, 2 = borderline, 3 = pass, 4 = high pass, 5 = excellent). The overall score was calculated as the mean of the item scores rounded to the nearest integer (1-5). There was space on the form for narrative feedback.

Examiners

At the start of the study the first author and the internal medicine education supervisor presented the ITA programme to a meeting of senior staff and residents in which the presentation was followed by a discussion of the programme. Before this meeting all the examiners had been sent a description of the test formats as well as the rating forms. Examiners received monthly schedules of the ITAs to be performed, except for the write-ups. The residents who assessed the write-ups were instructed to discuss at least three write-ups per week with each student they had to assess.

Population

We collected the completed rating forms from all 119 undergraduate students who entered the internal medicine clerkship at the Vrije Universiteit Medical Centre, Amsterdam between January 1999 and June 2002. The undergraduate medical curriculum of the VU Medical Centre offers four years of pre-clinical training followed by two years of clerkships in the major clinical disciplines. Before the start of the clerkship students take part in a six-week skills training programme, in which they practice skills on

manikins, on each other and on simulated patients. They then enter a three-week introductory clerkship in which they work with patients under strict supervision. The internal medicine clerkship is the first clinical rotation in which students participate in all tasks. Students spend the first six weeks of the rotation on the hospital wards and the last four weeks in the outpatient clinic. Every one to two weeks one student enters and another finishes the rotation.

Data analysis

An ITA programme is considered feasible when the majority of students obtain the required minimum number of assessments. We used two thirds of students obtaining 19 assessments as the arbitrary benchmark for good feasibility. Thus, we calculated the percentage of students obtaining the required number of assessments for the five ITA test formats. Additionally, we calculated the number of students obtaining the required number of assessments for each of the five test formats separately. Further data analysis was performed using only the data pertaining to the students who met the benchmark of 19 assessments.

In order to determine the reliability of the ITA programme, we calculated the mean scores and standard deviations for the 19 assessments and the five different test formats. A generalisability analysis was performed using the 19 assessments as items (items nested within students). This analysis probably overestimated reliability because of within-format dependence and large differences between formats in sampling frequency. Therefore, reliability was also estimated using the five different test formats as items (nested within students, because the content of the assessments differed between students) and averaging scores across assessments within each format. For the set of 19 assessments we also performed extrapolations to different numbers of items and calculated the corresponding generalisability coefficients. Extrapolation was not performed for the test formats. The reason for this was that format is a fixed facet, so that it is impossible to randomly add different formats to the existing ones.

Results

Of 119 students who entered the internal medicine clerkship during the study period, 106 (89%) completed the rotation and thus were included in the study.

Feasibility

A total of 73 students (69%) handed in completed forms for the required minimum number of assessments (i.e. they completed the required ITA programme). A total of 18 students (17%) failed to reach the benchmark for the clinical oral examinations (Table 1).

Table 1
The benchmark for feasibility, i.e. the required number of assessments per test format and the number and percentage of students reaching the benchmark

ITA	Benchmark	Number (%) of students meeting benchmark (n = 106)	
Student-patient encounter	1	102	(96%)
Critical appraisal session	1	102	(96%)
Case presentation	1	102	(96%)
Case write-up	12	104	(98%)
Structured long case	4	88	(83%)
Entire ITA programme	19	73	(69%)

Reliability

The mean score of the 73 students who completed the required ITA programme was 3.95 (SD 0.33) for the set of 19 assessments and 4.19 (SD 0.52) for the set of five test formats (Table 2). The mean scores for the 19 assessments ranged from 3.66 (0.60) for a case

write-up to 4.50 (0.63) for a structured long case. The mean scores for the five test formats ranged from 3.69 (0.55) for the student-patient encounter, to 4.30 (SD 0.49) for the structured long case. Reliability (generalisability coefficient) was 0.81 for the set of 19 assessments and 0.55 for the set of five test formats. Extrapolation from the set of 19 assessments revealed that 18-19 assessments were sufficient to reach a reliability of at least 0.80 (Table 3).

Table 2
Mean scores (standard deviations) by assessment and by test format for students meeting the benchmark of 19 assessments

ITA	Score for all assessments Mean (SD)	Score per test format Mean (SD)
Student-patient encounter	3.69 (.55)	3.69 (.55)
Critical appraisal session	3.91 (.65)	3.91 (.65)
Case presentation	4.16 (.60)	4.16 (.60)
Write-up	3.66 (.60) - 3.93 (.65)	3.85 (.37)
Structured long case	4.18 (.73) - 4.50 (.63)	4.30 (.49)
Overall	3.95 (.33)	4.19 (.52)

Table 3
Reliability (generalisability coefficient) as a function of the number of assessments

Number of assessments	Generalisability coefficient
10	0.70
15	0.77
19	0.81
30	0.87

Discussion

The aim of this study was to determine the feasibility and reliability of an ITA programme for the assessment of undergraduate students' clinical competence during a general internal medicine clerkship. The ITA programme consisted of four ward-based test formats (one student-patient encounter, one critical appraisal session, one case presentation, 12 case write-ups) and one outpatient-clinic based format (four structured long cases).

It was determined in advance that the ITA programme would be considered feasible when two thirds of the students met the required minimum number of assessments. The results provided evidence for the feasibility of the ITA programme. Nevertheless, 17% of the students failed to obtain assessments on four long cases in the outpatient clinic in the last week of the clerkship. Student-teacher interactions in the outpatient clinic are often described as sub-optimal because teachers face conflicting time pressures due to high patient turnover.[21] A recent study demonstrated the feasibility of integrating the educational activities of both students and teachers into the daily schedule of the outpatient clinic and scheduling protected time for educational purposes. These measures were found to be a way of optimising outpatient clinic teaching that was both feasible and satisfactory for students, staff and patients.[27] For this study, we integrated the student-patient encounter (the actual long case) and the time the student needed for the medical record and management plan into the daily activities of the outpatient clinic by explicitly scheduling students' time. In addition, the students had scheduled time for discussion with the member of staff. However, there was no similar provision for the 30 minutes of discussion and rating by staff, although the supervising staff member always saw the student and patient to supervise and round off the patient contact. However, no protected staff time was scheduled for 30 minutes of detailed discussion and rating, so that time pressure often interfered with this part of the long case. Setting aside time for this purpose in staff schedules might enhance the feasibility of the structured long cases and thus of the entire ITA programme.

The mean overall scores and reliability of the set of 19 assessments and the set of five test formats showed considerable differences. In the set of 19 assessments the relatively low mean scores for the 12 write-ups strongly affected the overall mean score. As a result the mean score of the set of 19 assessments was lower than that of the set of five test formats. The relatively low mean scores on the write-ups may be explained by the fact that write-ups are better integrated into the daily routine of clinical work than are the other aspects of ITA; writing and discussing write-ups is part of the daily work of both students and examiners (residents). This means that special preparation is almost impossible and examiners will tend to be critical. Reliability was sufficient for the set of 19 assessments, but probably overestimated. When reliability was estimated using 19 assessments as items, the case write-ups (12 out of 19 assessments) strongly affected the outcome and the same applies, albeit to a lesser extent, for the long cases. It would appear to be more appropriate to estimate reliability using the five test formats as items. When estimated in this way, reliability was somewhat disappointing compared with reliability based on 19 assessments. It was still superior, however, to reliability obtained for conventional clinical assessment formats. The overall reliability of the ITA programme might be improved by turning single-sample formats into multiple-sample ones. The precise numbers needed for the different test formats should be determined through further study using multivariate generalisability theory.[28]

There are two general drawbacks of this study. First, we focussed on the students that completed the clerkship. This meant that 11% of the students who entered the clerkship were excluded from the study. These students entered the clerkship in the study period but terminated the rotation prematurely and therefore did not finish the ITA programme. The percentage of drop-outs seems relatively high. However, students whose start of clerkship is delayed because they have failed the skills training programme are assigned to the internal medicine clerkship at the Vrije Universiteit Medical Centre. The majority of these students do not fail due to lack of ability, but becasue of chronic illness or other personal problems. When such

problems affect performance in the clerkship (for instance, if the student is unable to attend every day), students are advised to temporarily stop the clerkship and try to resolve their problems before resuming clerkship. The general impression of these students, when they were present, was that their performance did not differ markedly from that of the students who completed the ITA programme. This means that is unlikely that excluding the available test results of these students from the analysis affected the reliability estimates. Secondly, professional and social skills were only partly assessed by the ITA programme, which did not include competences such as work attitude and attitude towards colleagues. Turnbull *et al.* explored a method of ITA that assessed these competences and found very marginal reliabilities.[29] This warrants further exploration of methods to evaluate those competences.

Acknowledgements

We thank Professor A.J.M. Donker, emeritus professor of Internal Medicine, for enabling the implementation of the ITA programme in the Department of Internal Medicine and A. Thijs, educational co-ordinator, Department of Internal Medicine, for supporting the implementation of the programme.

References

1. Wass V, Vleuten van der C, Shatzer J, Jones R. Assessment of clinical competence. *The Lancet* 2001;357:945-9.
2. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990; 65:S63-7.
3. Magarian GJ, Mazur DJ. Evaluation of students in medicine clerkships. *Acad Med* 1990;65:341-5.
4. Kassebaum DG, Eagles RH. Shortcomings in the evaluation of students' clinical skills and behavior in medical school. *Acad Med* 1999;74:842-9.
5. Hull AL, Hodder S, Berger B, Ginsberg D, Lindheim N, Ouan J, Kleinhenz ME. Validity of three clinical performance assessments of internal medicine clerks. *Acad Med* 1995;70:517-22.
6. Harden RM, Gleeson FA. ASME Medical Education Booklet no.8. Assessment of medical competence using objective structured clinical examination (OSCE). *J Med Educ* 1979;13:41-54.
7. Mavis BE, Henry RC, Ogle KS, Hoppe RB. The emperor's new clothes: the OSCE reassessed. *Acad Med* 1996;71:447-53.
8. Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. *Ann Intern Med* 1995;123:795-9.
9. Turnbull J, van Barneveld C. Assessment of clinical performance: in-training evaluation. In: G.R. Norman, C.P.M. van der Vleuten & D.I. Newble, eds. *International handbook of research in medical education.* Dordrecht/Boston/ London: Kluwer Academic Publishers 2002;793-810.
10. Feletti G, Cameron D, Dawson-Saunders B, des Groseilliers JP, Dooley B, Farmer E, McAvoy P. In-training assessment. In: Newble D, Jolly B, Wakeford R, eds. *The certification and recertification of doctors: issues in the assessment of clinical competence.* Cambridge: Cambridge University Press 1994;151-66.
11. Gleeson F. The effect of immediate feedback on clinical skills using the OSLER. In: Rothman AI, Cohen R, eds. *Proceedings of the Sixth Ottawa Conference on Medical Education.* Toronto: University of Toronto Bookstore Custom Publishing 1994;412-5.
12. Wass V, Jones R, van der Vleuten C. Standardized or real patients to test clinical competence? The long case revisited. *Med Educ* 2001;35:321-5.
13. Daelmans HEM, Scherpbier AJJA, Vleuten van der CPM, Donker AJM. Reliability of clinical oral examinations re-examined. *Med Teach* 2001;23:422-4.
14. Metz JCM, Stoelinga GBA, Pels Rijcken-van Erp Taalman Kip EH, van den Brand-Valkenburg BWM. *Blueprint 1994: Training of doctors in the Netherlands. Objectives of undergraduate medical education in the Netherlands.* Nijmegen: University Publication Office 1994.

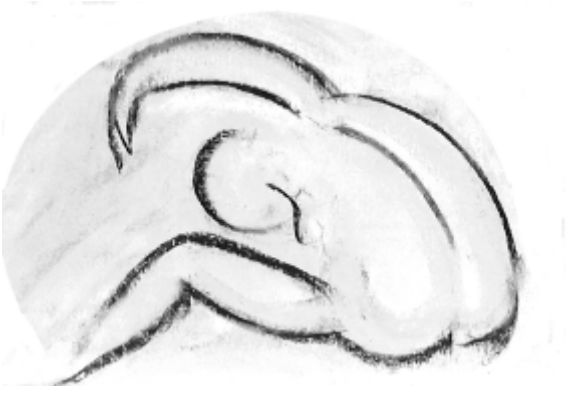
15. The Medical School Objectives Writing Group. Learning objectives for medical students education-guidelines for medical schools: Report I of the Medical School Objectives Project. *Acad Med* 1999;74:13-8.
16. Societal Needs Working Group, CanMEDS 2000 project. Skills for the new millennium. *Annales CRMCC* 1996;29:206-16.
17. Simpson JG, Furnace J, Crosby J, Cumming AD, Evans PA, Friedman Ben David M, Harden RM, Lloyd D, McKenzie H, McLachlan JC, McPhate GF, Percy-Robb IW, MacPherson SG. The Scottish doctor - learning outcomes for the medical undergraduate in Scotland: a foundation for competent and reflective practitioners. *Med Teach* 2002;24:136-43.
18. Van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ* 1996;1:41-67.
19. Schuwirth LW, Southgate L, Page GG, Paget NS, Lescop JM, Lew SR, Wade WB, Baron Maldonado M. When enough is enough: a conceptual basis for fair and defensible practice performance assessment. *Med Educ* 2002; 36:925-30.
20. Remmen R, Denekens J, Scherpbier AJJA, van der Vleuten CPM, Hermann I, Puymbroeck van H, Bossaert L. Evaluation of skills training during clerkships using student focus groups. *Med Teach* 1998;20:428-31.
21. Irby DM. Teaching and learning in ambulatory care settings: a thematic review of the literature. *Acad Med* 1995;70:898-931.
22. Norman GR. Objective measurement of clinical performance. *Med Educ* 1985;19:43-7.
23. Remmen R, Denekens J, Scherpbier A, Hermann I, van der Vleuten C, van Royen P, Bossaert L. An evaluation study of the didactic quality of clerkships. *Med Educ* 2000;34:460-4.
24. Hardy KJ, Demos LL, McNeil JJ. Undergraduate surgical examinations: an appraisal of the clinical orals. *Med Educ* 1998;32:582-9.
25. Meadow R. The structured exam has taken over. *BMJ* 1998;317:1329.
26. Wass V, Jolly B. Does observation add to the validity of the long case? *Med Educ* 2001;35:729-34.
27. Regan-Smith M, Young WW, Keller AM. An efficient and effective teaching model for ambulatory education. *Acad Med* 2002;77:593-9.
28. Hays RB, van der Vleuten, CPM, Fabb WE, Spike NA. Longitudinal reliability of the Royal Australian College of General Practitioners Certification Examination. *Med Educ* 1995;29:317-21.
29. Turnbull J, MacFadyen J, van Barneveld C, Norman G. Clinical work sampling: a new approach to the problem of in training evaluation. *J Gen Intern Med* 2000;15:556-61.

Chapter 4

Global clinical performance rating, reliability and validity in an undergraduate clerkship

HEM Daelmans, HH van der Hem-Stokroos, RJI Hoogenboom, AJJA Scherpbier, CDA Stehouwer, CPM van der Vleuten.

Accepted for publication in *The Netherlands Journal of Medicine*



Summary

Global performance rating is frequently used in clinical training despite its known psychometric drawbacks. Inter-rater reliability is low in undergraduate training and better in residency training, possibly because residency offers more opportunities for supervision. The predictive validity of global performance ratings in undergraduate and residency training is low to moderate, possibly due to low or unknown reliability of both global performance ratings and criterion measures. In an undergraduate clerkship, we investigated whether reliability improves when raters are more familiar with students' work and whether validity improves with increased reliability of predictor and criterion instrument.
Inter-rater reliability was determined in a clerkship with more student-rater contacts than usual. The in-training assessment programme of the immediately following clerkship was used as the criterion measure to determine predictive validity.
With four ratings, inter-rater reliability was 0.41 and predictive validity was 0.32. Reliability was lower and validity slightly higher than similar results published for residency training.
Even with increased student-rater interaction, the reliability and validity of global performance ratings were too low to warrant the usage of global performance ratings as individual assessment format. However, combined with other assessment measures, global performance ratings may lead to improved integral assessment.

Overview

What is already known

· Global performance ratings (GPR) show lower reliability in undergraduate compared with residency training.
· The predictive validity of GPRs for other measurements is low to moderate in undergraduate and residency training.

What this study adds

· The limited duration of clerkship rotations may diminish the effect of increased student rater contact on the reliability of GPRs.
· Improved estimation of the predictive validity of GPRs yields higher outcomes.
· GPRs can make a positive contribution to the evaluation of students' performance.

Suggestions for further research

· Further research might explore the impact of combining GPRs with other assessment methods.

Introduction

Evaluation of clinical performance typically takes the form of a global rating by a supervisor, halfway or at the end of a clinical rotation, covering learners' performance on a number of clinically relevant competences over a certain period of time. Hereafter we will refer to this type of rating as global performance rating (GPR). Despite the availability of new assessment methods, GPRs continue to be frequently used in both undergraduate and residency training most probably due to the combined advantage of feasibility and face validity (the assessed performance represents the performance domain of interest). In undergraduate training, GPRs are often the primary determinant of the final grade a student receives at the end of a clerkship.[1,2] Moreover, despite measures to increase the reliability of GPRs, such as rater training, in practice most assessors are not trained. At best the items on a scale are anchored to descriptors of criterion behaviour. In the last two decades several

studies have examined the reliability and validity of GPRs of untrained assessors in both undergraduate and residency training.

For inter-rater agreement among members of staff as a measure of the reliability of GPRs in undergraduate training the findings varied, with inter-rater agreement ranging from 0.29 to 0.42.[3,4] Studies performed in residency training have consistently demonstrated higher inter-rater agreement (0.79-0.87) than studies among undergraduate students.[5-8] A possible explanation for this difference may be that clinical staff, who typically evaluate students' and residents' performance, generally have more opportunity to supervise the work of residents than that of students because residency rotations last longer than clerkship rotations.[9,10] Moreover, because residents treat patients, supervision of residents is necessary to ensure the provision of appropriate patient care. Staff members have a strong professional stake in residents' adequate performance, because they may be held liable for adequate supervision.[11]

Assuming that the reliability of assessment benefits from increased supervision, we designed a study in which we measured inter-rater agreement on GPRs in a setting where staff members supervised students' work more frequently than is customary in undergraduate clerkships. If our assumption is correct, we would expect inter-rater agreement in this setting to be moderate to quite high.

Studies have investigated both concurrent and predictive validity of GPRs of untrained assessors. Both concurrent and predictive validity indicate the extent to which GPRs predict scores on a selected criterion that is not directly measured by the assessment but that is assumed to be parallel. For concurrent validity the criterion measurement is performed at the same time, for predictive validity it is performed at some point in the future. *Concurrent validity* has been studied in both undergraduate and residency training by correlating GPRs with more objective performance measures for the same training period, such as written examinations, OSCEs or simulated patient exams.[6,12-14] Correlations ranged from 0.19 to 0.33 for undergraduate training and from 0.29 to

0.56 for residency training. Fewer studies have addressed the *predictive validity* of GPRs. In undergraduate training predictive validity has been determined by comparing GPRs of student performance in different rotations with GPRs of end-of-clerkship performance or of performance in residency training. Predictive validity ranged from 0.17 to 0.44.[12,15] Callahan examined the predictive validity of GPRs in clerkships for the results on United States Medical Licensing Examinations (USMLE) steps 2 and 3. The maximum predictive validity was 0.29 for USMLE step 2 and for USMLE step 3 it was 0.20.[15] In residency training the predictive validity of GPRs for performance at in-training and American Board of Internal Medicine (ABIM) certification exams was reported to be moderate for overall competence (0.19) and for specific competences on a global performance rating scale (ranging from 0.11 to 0.41).[8,16] The validity coefficients reported in these studies suffered from attenuation, i.e. low or unknown reliabilities in predictor and criterion variables introduce inaccuracy into the calculation. When measurement error is present in one or both of the variables that are being correlated, the correlation coefficient that is obtained will be attenuated. This implies that the observed correlation coefficient between less than perfectly reliable scores will tend to underestimate the true level of co-variation between the predictor and criterion variables.[17] Therefore, if the reliability of either the predictor or the criterion variables is low, validity coefficients will also be low. This effect might have been even stronger in studies in undergraduate training settings, where the reliability of global ratings was typically low and the reliability of the criterion variable mostly unknown. That is why we considered it worthwhile to examine the predictive validity of GPRs in undergraduate training using a criterion variable of known and acceptable reliability. If staff members have more opportunities to supervise students' performance the reliability of GPRs in undergraduate medical training might benefit.

We thus sought to answer the following research questions: What is the reliability of GPRs in an undergraduate clerkship with increased rater-student interactions? And, because of the inaccuracy in

reliability estimates of GPRs due to the low or unknown reliabilities of predictor and criterion variables (attenuation): what is the validity of GPRs when the reliabilities of both predictor and criterion variables are assumed to be perfect (disattenuation correction)?

We addressed the research questions by determining the inter-rater agreement for GPRs in an undergraduate clerkship with extensive interaction (detailed below) of staff members (raters) and students. These GPRs were then compared with a valid and reliable performance measure for the competences demonstrated by the same students in the immediately following clerkship with less student staff interaction. The performance measure used in the second rotation was the overall score on an in-training assessment programme (ITA) consisting of several assessments of clinical competence.

Materials and Methods

Educational background

At the Vrije Universiteit Medical Centre (VU medical centre), Amsterdam, the Netherlands, four years of preclinical medical education are followed by two years of rotations in the major clinical disciplines. The clinical phase starts in Year 5 with a three-week introductory clerkship in which the students are closely supervised by clinical staff. The students' main tasks are history taking and physical examination, medical record writing and practising skills in pathophysiological thinking and clinical reasoning in structured discussions both in groups supervised by a member of staff and in writing. Staff members are scheduled to supervise the group discussion, discuss the medical records and observe (parts of) history taking and physical examination. Every day a different staff member supervises the students in their daily structured group discussion, which lasts about an hour. The supervisor asks several students to elaborate on their findings, interpret data, formulate differential diagnoses and propose additional

investigations. Over the course of the clerkship, students are supervised twice by a member of staff while performing a (scheduled) complete patient interview and physical examination. Afterwards student and staff member discuss the student's performance. For morning and evening reports, radiology meetings, interdisciplinary meetings et cetera, students are not linked to residents but to members of staff, who are thus more focussed on students' contributions. In most cases, a student is supervised by six to seven members of staff during the three-week rotation. At the end of the three weeks, a staff member to whom this task has been assigned determines a GPR for the student's performance during the rotation. Next, the students move on to the ten-week Internal Medicine rotation in the VU medical centre or in one of the affiliated hospitals. This rotation is scheduled immediately following the introductory clerkship. In this rotation the students are mostly supervised by residents instead of members of staff during student-patient interactions and for medical records. This rotation involves more participation by the students in day-to-day clinical practice, including multidisciplinary meetings and on call duties. In order to better monitor students' performance in the Internal Medicine clerkship, a programme of systematic observation and documentation of students' actual performance (detailed below) has been introduced in the rotation in the VU medical centre.

Global clinical performance rating (GPR)

Eight supervisors of the introductory clinical rotation were approached during a staff meeting and asked to participate in the study. Participation entailed giving a global performance rating on a five-point scale (1 = fail, 2 = borderline, 3 = pass, 4 = high pass, 5 = excellent) for every student doing the rotation in the study period. The raters received a brief description to be used in rating students' performance. The description mainly focused on the comparison of the performance of the student to that of an average student in his/her first three weeks of clinical rotation. On a student's last day of

this rotation, the members of staff received a form together with a scanned picture of the student concerned. The members of staff that had interacted with the student were asked to complete and return the form. The members of staff that did not supervise the student could return the form without filling it out. The entire procedure was computerised. We used a single-item rating (global performance) to preclude the use by raters of only one or two items (dimensions of performance) of a larger scale to judge global performance.[5,6,16,18,19] Each participating staff member was asked to complete one evaluation form for each student during the study period. In this way students could receive a maximum of eight GPRs from different examiners.

In-training assessment

All students proceed from the introductory rotation to the Internal Medicine rotation. In the Internal Medicine rotation in the VU medical center a fully integrated in training assessment (ITA)-programme is used. ITA implies systematic observation and documentation of learners' actual performance using several formats.[20] ITA in undergraduate clinical training has been described as a feasible assessment format that has reasonable reliability and good content validity.[20-22] The ITA programme used in this study consisted of observation and documentation of students' actual performance in five test formats.[22] A minimum frequency per student over the entire clerkship was specified for each test format, resulting in a required total of nineteen assessments: three single-sample formats (student-patient encounter, critical appraisal session and case presentation) and two multiple-sample formats (twelve case write-ups and four structured long cases). The student-patient encounter, critical appraisal session, case presentation and structured long cases were assessed by staff and the case write-ups by residents. All tests were rated on a five-point Likert scale (1 = fail, 2 = borderline, 3 = pass, 4 = high pass, 5 = excellent) and an overall score was obtained by calculating the mean of the scores and rounding it to the nearest

integer (1-5). The assessors of the ITA programme were not specifically informed about the current study.

Subjects: student participants

From April 2001 until October 2002, 91 students received global ratings of their performance in the introductory clerkship. We collected ITA scores for 48 of these 91 students. These 48 students did the subsequent ten week Internal Medicine rotation in the VU medical center, whereas the other students went to affiliated hospitals, where the ITA programme had not yet been implemented. A t-test on the means of the GPRs showed no differences between the GPRs of the students participating in the study and the students assigned to the affiliated hospitals.

Data analysis

First we counted the total number of GPRs per student. For further calculations we used the balanced data set of the group of students for whom at least four GPRs were available.[23] For the analysis we used random samples of four GPRs per student.

We calculated means and standard deviations for the GPRs. Inter-rater reliability was estimated based on generalisability theory. We used a one-facet design with raters (or GPRs) nested within persons (students) to estimate variance components. Subsequently, reliability coefficients, i.e. dependability coefficients, were calculated as a function of the number of examiners (or GPRs). In the clerkship studied, each member of staff supervised students in two to three group discussions and most probably several times during reports and meetings. Only two members of staff witnessed complete student-patient contacts (interview and physical examination). As a result, these two staff members may have developed significantly different judgements than did other staff members. However, the fact that different selections of four members of staff did not yield significantly different ratings suggests that this was not the case.

We calculated means and standard deviations of the ITA scores. A similar generalisability design was used to estimate the reliability of the ITA programme, with observations across test formats nested within students.

The predictive validity of the GPRs was determined by correlating the mean GPR with the mean ITA score. We estimated the disattenuated correlation (the estimated correlation when both predictor and criterion measures have perfect reliability) using the reliability coefficient of the GPRs with four raters and the reliability of the ITA programme.

Table 1
The number of global performance ratings (GPRs) per student

Number of GPRs	2	3	4	5	6	7	8
Number of students	1	3	13	30	16	23	5

Results

Of the 91 students whose performance was rated in the introductory clerkship, 87 received four or more GPRs (Table 1). Four students were rated by fewer than 4 staff members. Each of the 8 members of staff who participated in the study contributed to the GPRs throughout the duration of the study. Having had no interaction with the student, holidays and illness were the main reasons given by staff members for not having witnessed a student's performance and thus being unable to provide a GPR.

The mean GPR was 3.19 (SD 0.37). Inter-rater reliability with four GPRs per student was 0.41 (n=87). Twenty-five GPRs per student would be needed to reach sufficient reliability (0.8) (Table 2).

Table 2
Reliability coefficients as a function of the number of examiners or global performance ratings (GPRs)

Number of GPRs	4	6	10	25
Reliability coefficient	0.41	0.51	0.63	0.81

Means and standard deviations of the different ITAs ranged from 3.61 (SD 0.65) for case write-ups to 4.35 (SD 0.65) for case presentations (Table 3).
The reliability of the ITA programme was 0.71. The observed predictive validity of the GPRs for the ITA programme was 0.32 ($p<0.05$) and the disattenuated predictive validity was 0.59.

Table 3
Mean scores (standard deviations) for the different tests in the ITA programme

ITA	**Mean (SD)**
Student-patient encounter	3.70 (.70)
Critical appraisal session	4.00 (.67)
Case presentation	4.35 (.65)
Write-up	3.61 (.65) - 4.04 (.71)
Structured long case	4.00 (.67) - 4.26 (.63)

Discussion

Global ratings have some well-known disadvantages. They are often only given at the end of a rotation when assessors may have forgotten details of students' performance. In addition, they may be biased due to a halo effect, i.e. the phenomenon that an impression created by a student's good or poor performance in one area affects

assessors' judgements of student's performance in another area.[24] In the introductory clerkship, we could easily have used structured assessment with rating forms, such as in-training assessment. However, we purposely used global ratings, because in this study we set out to investigate the possibility of improving the reliability and validity of such ratings, as they continue to be much used in undergraduate and graduate training. With improved reliability and validity, global ratings could make a truly positive contribution to assessment of clinical performance, the more so since they can cover more competences than assessment formats focused on specific items.[25]

We investigated the reliability and the predictive validity of GPRs in undergraduate training. We studied the reliability of GPRs in an introductory clerkship where the members of staff who rated the students supervised students' performance more frequently than is customary in undergraduate clerkships. We expected that this would yield a better, i.e. moderate to high, reliability than is generally reported for GPRs in undergraduate clinical training. We observed an inter-rater reliability of 0.41, which is comparable with the literature on undergraduate inter-rater agreement. We speculate that the potentially positive influence of increased supervision of students' clinical work by staff may have been mitigated by the limited duration of the clerkship as compared to residency rotations.[26] A relatively shorter period during which staff are in a position to supervise students may lead to a correspondingly diminished accuracy of the perceived levels of students' performance. Hence increased supervision did not result in improved reliability. Furthermore, reliability may have (slightly) suffered on account of staff not having participated in assessment training before this study was conducted.[27,28] The results showed that 25 GPRs from different examiners would be needed to achieve adequate reliability. Other studies have yielded estimates of between 7 and 14 ratings to attain a reliability of 0.80.[5,29,30] However, the assessment formats on which the GPRs in those studies were based included aspects that might potentially improve reliability, such as a long duration of the student-staff work relationship (up to one year),

a highly detailed description of the behaviour associated with the low and high scale points on the rating scale and raters who were better acquainted with students' performance (e.g. resident ratings). We suspect that more ratings will be needed to reach acceptable reliability in undergraduate settings, where the working relationship of staff and students lasts only a short time and raters thus witness less of the students' work and have to judge performance without guidance from concrete descriptions of the behaviours corresponding to the different scale points.

The validity measure derived from the predictive validity of the GPRs for the scores on the ITA programme in the subsequent clerkship was 0.32. Although slightly higher than the predictive validity reported for overall competence in studies in both undergraduate and residency training, it is still quite low.[8,15] The GPRs in this study were based on staff members' evaluations of students at the end of a three-week rotation. Despite frequent student-staff interaction in these three weeks, details of the interactions can be lost quite quickly.[31,32] By contrast, the evaluations in the ITA programme were recorded immediately following the activity or behaviour that was evaluated and according to a checklist. Despite the more than usually intensive interaction between staff and students in the initial clerkship, the fact that the GPRs were based on less detailed information about students' clinical performance than the ITA scores may offer an explanation for the low predictive validity of the GPRs. Disattenuated predictive validity was 0.59, however, which is much higher. Our findings implicate that GPRs, despite being based on less detailed information, can still make a positive contribution to the evaluation of students' performance. In a recently published study, Kreiter and Ferguson found comparable disattenuated predictive validity when they compared global ratings of clinical clerkship performance with former measures of physical examination performance provided by simulated patients (SP) using ratings and checklists, and with SP ratings of rapport and communication.[33] They conclude that measures of skills by global ratings are correlated with other clinical performance measures and

discuss that more studies of this topic are needed to conclude that global ratings make a positive contribution to students' evaluation. The evidence in our study points in the same direction and thus contributes to the conclusion that global ratings can positively contribute to students' evaluation.

This study has one major drawback. We compared our findings to findings in the literature and not to those of a control group. The circumstances in which research in the presented literature was performed were certainly different from the circumstances of our study. However, it was practically not feasible to have a control group in the same clerkship at the same time due to staff shortage and the practical impossibility to have two educational programmes performed by the same members of staff during the same period of time.

Our results indicate that even when conditions in an undergraduate rotation are positively manipulated, reliability and validity of GPRs remain low. However, the reliability and validity we reached were not lower than those found for other assessment formats performed over a short testing time.[34,35] This means that GPRs can contribute to the assessment of undergraduate students' clinical competences as long as they are sampled on many occasions and by many assessors. Nevertheless, sufficient reliability and validity are likely to be hard to achieve. In a recent review, Williams et al. concluded that GPRs by themselves were an insufficient measure of students' clinical competence, even though they might be an important source of information about it.[36] These authors recommended that GPRs should be supplemented with ratings of students' performance in standardized clinical encounters and assessment protocols. The results of our study point to a similar recommendation, i.e. to combine GPRs with more specific and reliable assessment formats, such as the ITA programme in this study, to arrive at an integrated assessment programme. Further studies will have to examine whether such an assessment programme can provide reliable and valid measures of students' competences.

References

1. Magarian GJ, Mazur DJ. Evaluation of students in medicine clerkships. *Acad Med* 1990;65:341-5.
2. Kassebaum DG, Eaglen RH. Shortcomings in the evaluation of students' clinical skills and behavior in medical school. *Acad Med* 1999;74:842-9.
3. Dielman TE, Hull A, Davis WK. Psychometric properties of clinical performance ratings. *Eval Health Prof* 1980;3:103-17.
4. Maxim BR, Dielman TE. Dimensionality, internal consistency and inter-rater reliability of clinical performance ratings. *Med Educ* 1987;21:130-7.
5. Haber RJ, Avins AL. Do ratings on the American Board of Internal Medicine resident evaluation form detect differences in clinical competence. *J Gen Int Med* 1994;9:140-5.
6. Kwolek CJ, Donnelly MB, Sloan DA, Birrell SN, Strodel WE, Schwartz RW. Ward evaluations: should they be abandoned? *J Surg Res* 1997;69:1-6.
7. Davis JD. Comparison of faculty, peer, self and nurse assessment of obstetrics and gynecology residents. *Obstet Gynecol* 2002;99:647-51.
8. Durning SJ, Cation LJ, Jackson JL. The reliability and validity of the American Board of Internal Medicine monthly evaluation form. *Acad Med* 2003;78:1175-82.
9. Remmen R, Denekens J, Scherpbier A, Hermann I, Van der Vleuten C, Royen PV, Bossaert L. An evaluation study of the didactic quality of clerkships. *Med Educ* 2000;34:460-4.
10. Busari JO, Scherpbier AJJA, Van der Vleuten CPM, Essed GGM. The perceptions of attending doctors of the role of residents as teachers of undergraduate clinical students. *Med Educ* 2003;37:241-7.
11. Kachalia A, Studdert DM. Professional liability issues in graduate medical education. *JAMA* 2004;292:1060-1
12. Keynan A, Friedman M, Benbassat J. Reliability of global rating scales in the assessment of clinical competence of medical students. *Med Educ* 1987,21:477-81.
13. DaRosa DA, Dawson-Saunders B, Folse R. A comparison of objective and subjective measures of clinical competence. *Eval Program Plann* 1985; 8:327-30.
14. Adusumilli S, Cohan RH, Korobkin M, Fitzgerald JT, Oh MS. Correlation between radiology resident rotation performance and examination scores. *Acad Radiol* 2000;7:920-6.
15. Callahan CA, Erdmann JB, Hojat M, Veloski JJ, Rattner S, Nasca TJ, Gonnella JS. Validity of faculty ratings of students' clinical competence in core clerkships in relation to scores on licensing examinations and supervisors' ratings in residency. *Acad Med* 2000;75:S71-3.

16. Norcini JJ, Webster GD, Grosso LJ, Blank LL, Benson JAJr. Ratings of residents' clinical competence and performance on certification examination. *J Med Educ* 1987;62:457-62.
17. Pedhazur EJ. *Multiple Regression in Behavioral Research - Explanation and Prediction*. 2nd Edition. New York: Holt, Rinehart and Winston eds. 1982;112-4.
18. Metheny WP. Limitations of physician ratings in the assessment of student clinical performance in an obstetrics and gynecology clerkship. *Obstet Gynecol* 1991;78:136-41.
19. Streiner DL. Global rating scales. In: Neufield VR, Norman GR, eds. *Assessing Clinical Competence*. New York: Springer Publishing Company 1985;114-41.
20. Turnbull J, Van Barneveld C. Assessment of clinical performance: in-training evaluation. In: Norman GR, Van der Vleuten CPM, Newble DI, eds. *International Handbook of Research in Medical Education.* Dordrecht/ Boston/London: Kluwer Academic Publishers 2002;793-810.
21. Turnbull J, MacFadyen J, Van Barneveld C, Norman G. Clinical work sampling: a new approach to the problem of in training evaluation. *J Gen Intern Med* 2000;15:556-61.
22. Daelmans HEM, Van der Hem-Stokroos HH, Hoogenboom R, Scherpbier AJJA, Stehouwer CDA, Van der Vleuten CPM. Feasibility and reliability of an in-training assessment programme in an undergraduate clerkship. *Med Educ* 2004;12:1270-7.
23. Kreiter DC, Ferguson K, Lee W, Brennan RL, Densen P. A generalizability study of a new standardized rating form used to evaluate students' clinical clerkship performances. *Acad Med* 1998;73:1294-8.
24. McKinstry BH, Cameron HS, Elton RA, Riley SC. Leniency and halo effects in marking undergraduate short research projects. *BMC Med Educ* 2004;4:28.
25. Van Luijk SJ, Van der Vleuten CPM. A comparison of checklists and rating scales in performance-based testing. In: Hart IR, Harden RM, DesMarchais J, eds. *Current Developments in Assessing Clinical Competence.* Montreal: Can Health Publications 1992;357-82.
26. Rothstein HR. Interrater reliability of job performance ratings: growth to asymptote level with increasing opportunity to observe. *J Appl Psychol* 1990;75:322-7.
27. Noel GL, Herbers JEJ, Caplow MP, Cooper GS, Pangaro LN, Harvey J. How well do internal medicine faculty members evaluate the clinical skills of residents? *Ann Int Med* 1992;80:1294-8.
28. Newble DI, Hoare J, Sheldrake PK. The selection and training of examiners for clinical examinations. *Med Educ* 1980;14:345-9.
29. Carline JD, Paauw DS, Thiede KW, Ramsey PG. Factors affecting the reliability of ratings of students' clinical skills in a medicine clerkship. *J Gen Int Med* 1992;7:506-10.

30. Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. *JAMA* 1993;269:1655-60.
31. Heneman RL. The effects of time delay in rating and amount of information observed on performance rating accuracy. *AMJ* 1983;26:677-86.
32. Kassin S, Tubb V, Hosch H, Memon A. On the "general acceptance" of eyewitness testimony research: a new survey of the experts. *Am Psychol* 2001;56:405-16.
33. Ferguson KJ, Kreiter CD. Using a longitudinal database to assess the validity of preceptors' ratings of clerkship performance. *Adv Health Sci Educ Theory Pract* 2004;9:39-46.
34. Swanson DB. A measurement framework for performance-based tests. In: Hart I, Harden RJ, eds. *Further Developments in Assessing Clinical Competence*. Montreal: Can-Health Publications 1987;13-45.
35. Petrusa ER. Clinical performance assessment. In: Norman GR, Van der Vleuten CPM, Newble DI, eds. *International Handbook of Research in Medical Education.* Dordrecht/Boston/London: Kluwer Academic Publishers 2002;673-709.
36. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teach Learn Med* 2003;15:270-92.

Chapter 5

Effectiveness of clinical rotations as a learning environment for achieving competences

HEM Daelmans, RJI Hoogenboom, AJM Donker,
AJJA Scherpbier, CDA Stehouwer, CPM van der Vleuten

Published in *Medical Teacher 2004;26:305-312*

Summary

Competences are becoming more and more prominent in undergraduate medical education. Workplace learning is regarded as crucial in competence learning. Assuming that effective learning depends on adequate supervision, feedback and assessment, we studied the occurrence of these three variables in relation to a set of clinical competences. We surveyed students at the end of their rotation in surgery, internal medicine or paediatrics asking them to indicate for each competence how often they had received observed and unobserved supervision, the seniority of the person who provided most of their feedback, and whether the competence was addressed in formal assessments. Supervision was found to be scarce and mostly unobserved. Senior staff did not provide much feedback, and assessment mostly targeted patient-related competences. For all variables, the variation between students exceeded that between disciplines. We conclude that conditions for adequate workplace learning are poorly met and that clerkship experiences show huge inter-student variation.

Overview

Adequate supervision, feedback and assessment are a necessary condition for effectively achieving clinical competences in the workplace setting.

Supervision of students performing clinical competences is rare and mostly unobserved.

Most feedback is given by persons with limited seniority.

Assessment mainly focuses on directly patient-based competences and on team working skills.

For all three variables individual variations between students are high.

Introduction

The goals of medical education are increasingly being defined in terms of competences rather than discrete learning objectives. Competences require integration of relevant knowledge, skills and attitudes to enable handling of complex situations and problems in an appropriate manner. When curricular goals are defined in terms of competences, the focus in education and assessment will have to be on the integration of skills, knowledge and attitudes rather than on isolated (components of) skills and knowledge. This will further authentic learning and help bridge the gap between theory and practice. As evidenced by reports from North America, Europe and Australia, the shift from detailed learning objectives to integrated competences is an international phenomenon.[1-8]

Because of their integrated nature, competences are best learned in an authentic learning environment. Thus workplace learning is of vital importance. From this perspective it becomes important to examine the quality of the workplace as a learning environment. A number of studies have indicated that the effectiveness of learning in the workplace is not always optimal.[9-12] Kilminster *et al.* (2002) found supervision to be an essential factor and others have pointed out that feedback and assessment play a major role in workplace learning.[13-16] Unfortunately, these three factors occur rather infrequently, with feedback often being provided by professionals that are not fully qualified and assessment lacking sufficient congruence with the intended objectives.[17,18] Most of the studies referred to focus on separate procedural and clinical skills and did not address competences as learning goals. We wanted to investigate the effectiveness of the workplace as a learning environment for achieving competences. Our investigation focused on clerkship learning, because clinical rotations are the prominent form of workplace learning in the undergraduate medical curriculum. Our study is based on the view that adequate and appropriate supervision, feedback and assessment are prerequisites for effective workplace learning. We conducted a survey to obtain students' views on the frequency of supervision,

feedback and assessment in relation to a set of specified competences (detailed below). In addition we compared the frequencies across clerkships in different disciplines.

Methods

Population and educational context

Over a six-month period we asked 104 undergraduate medical students of the Vrije Universiteit Medical Centre in Amsterdam to complete a questionnaire. The students had just finished a clinical rotation in the Department of Internal Medicine (28 students), Surgery (40 students) or Paediatrics (36 students). The sample size within departments was considered sufficient to allow reliable inferences.[19] The Vrije Universiteit Medical Centre offers an interdisciplinary, mainly lecture-based curriculum. Four pre-clinical years are followed by two years of clinical clerkships. Prior to the clinical phase students receive skills training in a skills laboratory. The emphasis of the skills programme is on procedural and communication skills. The first clerkship attachment is in internal medicine. The order of subsequent rotations varies. The clerkships are discipline-based and conventional in educational set-up.

Instrument

We developed a questionnaire for this study asking students to indicate whether they had received supervision, feedback and had been assessed in relation to a set of 16 competences. We included observed and unobserved supervision, the latter being either findings checked or findings discussed at a later time than the actual students' performance.

The frequency of observed and unobserved *supervision* was examined for seven settings in which supervision was likely to occur: student-patient contacts, review of medical records, ward rounds, paper rounds, case presentations, critical appraisal sessions

(presentation and discussion of scientific papers) and bedside teaching (see Table 1). For these seven educational settings students were asked to indicate how often they had been supervised when performing the various competences. Students were asked to indicate the frequency of both observed and unobserved supervision on a three-point scale: on less than 35% of all occasions, at least 35% but less than 70% of all occasions, and at least 70% or more of all occasions. The figure of 35% was chosen as an arbitrary minimum benchmark of supervision and 70% was considered to be a perfect score.

Students were asked to indicate which of three groups of potential feedback providers actually gave most of the *feedback*: 1) peers, i.e. other students as well as nursing and paramedical staff, 2) residents and 3) academic clinical staff. The quality of the feedback was inferred from the seniority of the person who gave it.[20] Students were asked to indicate for each competence if it was addressed in the regular *assessment* programme, i.e. the intermediate progress interview, the final interview and the unobserved clinical oral examination (all unstructured). They could choose from the answer options: yes, no or not applicable. The latter option was to be used when the question was not relevant in relation to a particular assessment.

The competences included in the questionnaire were derived from 'Blueprint 1994, training of doctors in the Netherlands'.[4] This document is the result of a consensus procedure among experts. It contains non-disciplinary competences and discipline based clinical and procedural skills and problems as starting points for training. A panel of three clinical education co-ordinators selected the competences they considered most relevant for this study from the non-disciplinary competences for internal medicine, paediatrics and surgery. Sixteen out of twenty competences were included (Table 1). This set of competences is very similar to competence or outcome lists used in other countries.[1,5-7]

Procedure

All students were asked to complete the questionnaire after their last assessment in the clerkship rotation they had just completed. Because the clinical oral examination for internal medicine is scheduled two years after the rotation in internal medicine, no data on that assessment were obtained.

Statistical analysis

Mean frequencies were calculated across students by competence and by discipline. For *supervision* we calculated the mean percentage of respondents who reported supervision of a competence in less than 35% or more than 70% of the selected educational occasions. For *feedback* we calculated the mean percentages of students indicating which category of supervisors provided most of the feedback for each of the different competences. For *assessment* we calculated for each competence the percentage of students who reported that it was addressed in a particular formal evaluation. We used the SPSS programme (version 11) for all calculations.

Results

Population

Eighty-one students returned the questionnaires: 25 from internal medicine, 28 from surgery and 28 from paediatrics. The overall response rate was 78% (internal medicine 89%, surgery 70% and paediatrics 78%). Non response was mainly due to organisational shortcomings.

Supervision

In Table 2 the mean percentages of students (with standard deviations) reporting a particular frequency of supervision are shown by discipline and competence.

Table 1

Overview of selected competences and educational events during which the competences can be supervised

	Educational setting						
Competences	Student-patient contact	Review of medical record	Ward round	Paper round	Case Presen-tation	Critical appraisal session	Bedside teaching
1 Listing problems and requests for help		+	++	+	+		++
2 Taking a history	++	+			+		++
3 General physical examination	++	+	++		+		++
4 Interpreting and evaluating data [a]		+		+	+		
5 Formulating a differential diagnosis [b]		+		+	+		
6 Proposing additional investigations		+		+	+		
7 Outcome of additional investigations [c]		+	+	+	+		+
8 Making a management plan		+		+	+		
9 Evaluating the result of treatment	++		++	+	+		+
10 Giving information to the patient	++						
11 Writing a letter of referral, or a request for consultation or an additional investigation		+		+			
12 Writing medical records in accordance with legislation		+		+	+		+
13 Writing a letter of discharge		+					
14 Maintaining professional competence[d]		+		+	+	++	
15 Building a doctor-patient relationship	++	+	++	+			++
16 Working in a team			++	+			

+ unobserved; ++ observed; [a] data from problem description, history and physical examination; [b] differential diagnosis or problem list, probability diagnosis or working hypothesis; [c] interpreting and evaluating the outcome; [d] advancing and maintaining professional competence through actively searching for relevant literature, reviewing professional literature critically and taking responsibility for one's own lifelong learning

Across the three disciplines observed supervision was sparse. For unobserved supervision the 70% benchmark was reported more often but still at a limited rate, ranging from 0% to 33% across competences. For all competences, a sizeable group of students, ranging from 23 to 98%, reported that supervision failed to reach the 35% benchmark. Supervision was reported to occur most frequently for patient-related competences. Most of the supervision occurred without observation. The huge standard deviations are indicative of strong variation among students. Reanalysis using median values yielded no differences of interpretation. Supervision was reported to occur most frequently for internal medicine and least frequently for paediatrics. Differences between individual students far exceeded the differences between disciplines.

Feedback

Table 3 presents the mean percentages and standard deviations of students indicating a particular group as providing most of their feedback.

The results show that most of the feedback was provided by residents. Academic clinical staff provided only a limited amount of feedback, except for the competence 'maintaining professional competence'. This finding is probably due to mandatory clinical staff attendance at critical appraisal sessions. Peers and paramedical staff provide hardly any feedback. As was the case for supervision, standard deviations were large, pointing to strong individual variation among students. Across clerkships most staff involvement was seen in internal medicine.

Assessment

The results for assessment are given in Table 4.

The lowest percentages were found for competences related to written and verbal communication. The highest frequencies were found for directly patient-related competences. In all disciplines, team working skills were a frequent feature in assessment. For all

Table 2

Mean frequency (and standard deviation) of students who reported competences as being supervised less than 35% and more than 70% of the selected educational events, in percentages per competence

	Internal Medicine				Surgery				Paediatrics			
	unobserved		observed		unobserved		observed		unobserved		observed	
Competence	**<35%**	**>70%**	**<35%**	**>70%**	**<35%**	**>70%**	**<35%**	**>70%**	**<35%**	**>70%**	**<35%**	**>70%**
1 Listing problems and request for help	49 (35)	20 (32)	68 (35)	10 (17)	61 (39)	12 (26)	73 (27)	12 (21)	60 (33)	11 (18)	92 (18)	0
2 Taking a history	34 (37)	24 (36)	76 (26)	11 (16)	39 (39)	16 (31)	68 (25)	11 (18)	59 (39)	16 (27)	77 (23)	11 (17)
3 Performing a general physical examination	44 (42)	20 (38)	65 (31)	13 (22)	48 (40)	13 (29)	73 (24)	9 (14)	59 (41)	9 (20)	80 (22)	5 (10)
4 Interpreting and evaluating data	23 (34)	33 (43)	-	-	50 (29)	13 (26)	-	-	62 (36)	14 (21)	-	-
5 Formulating a differential diagnosis	31 (37)	27 (35)	-	-	52 (32)	17 (27)	-	-	62 (35)	19 (22)	-	-
6 Proposing additional investigations	35 (34)	13 (29)	-	-	49 (32)	10 (24)	-	-	64 (35)	18 (32)	-	-
7 Outcome of additional investigations	60 (31)	10 (21)	-	-	56 (32)	8 (20)	-	-	81 (29)	6 (22)	-	-
8 Making a management plan	45 (41)	11 (28)	-	-	46 (31)	18 (32)	-	-	71 (31)	12 (28)	-	-
9 Evaluating the result of treatment	65 (39)	12 (27)	78 (41)	6 (22)	68 (36)	11 (24)	61 (42)	19 (37)	82 (33)	0	86 (30)	0
10 Giving information to the patient	-	-	88 (34)	4 (20)	-	-	82 (39)	7 (26)	-	-	92 (27)	0
11 Writing a letter of referral, or a request for consultation or an additional investigation	72 (41)	10 (25)		-	81 (28)	4 (13)	-	-	86 (30)	4 (13)	-	-
12 Writing medical records in accordance with legislation	81 (29)	7 (21)		-	83 (28)	6 (18)	-	-	92 (22)	3 (14)	-	-
13 Writing a letter of discharge	64 (49)	20 (41)		-	79 (42)	4 (19)	-	-	82 (39)	7 (26)	-	-
14 Maintaining professional competence	51 (44)	14 (27)	33 (48)	33 (48)	78 (26)	6 (13)	36 (49)	25 (44)	79 (22)	7 (11)	38 (50)	23 (43)
15 Building a doctor-patient relationship	86 (27)	0	73 (32)	3 (9)	95 (21)	0	85 (29)	6 (16)	98 (10)	0	97 (11)	0
16 Working in a team	83 (38)	4 (20)	50 (51)	14 (35)	79 (42)	4 (19)	64 (49)	25 (44)	96 (19)	0	89 (32)	0

Table 3

Mean frequency (and standard deviation) of students who reported supervisor categories to be the most involved in feedback of competences, in percentages per competence

Competence	Internal Medicine			Surgery			Paediatrics		
	peers*	resident	staff	peers*	resident	staff	peers*	resident	staff
1 Listing problems and requests for help	4 (20)	76 (44)	20 (41)	7 (26)	86 (36)	7 (26)	7 (27)	85 (36)	8 (27)
2 Taking a history	14 (35)	77 (43)	9 (29)	4 (20)	96 (20)	0	4 (20)	85 (37)	11 (33)
3 Performing a general physical examination	5 (21)	68 (48)	27 (46)	7 (27)	93 (27)	0	7 (27)	93 (27)	0
4 Interpreting and evaluating data	0	70 (47)	30 (47)	0	96 (19)	4 (19)	0	88 (33)	12 (33)
5 Formulating a differential diagnosis	0	87 (34)	13 (34)	0	93 (27)	7 (27)	0	80 (41)	20 (41)
6 Proposing additional investigations	0	82 (39)	18 (39)	0	93 (26)	7 (26)	0	85 (37)	15 (37)
7 Outcome of additional investigations	0	91 (29)	9 (29)	0	88 (33)	12 (33)	0	81 (40)	19 (40)
8 Making a management plan	0	74 (45)	26 (45)	0	89 (32)	11 (32)	0	88 (33)	12 (33)
9 Evaluating the result of treatment	0	80 (41)	20 (41)	4 (20)	84 (37)	12 (33)	8 (28)	60 (50)	32 (48)
10 Giving information to the patient	18 (39)	68 (48)	14 (35)	8 (28)	84 (38)	8 (28)	10 (30)	71 (46)	19 (40)
11 Writing a letter of referral, or a request for consultation or an additional investigations	0	88 (34)	12 (34)	0	92 (27)	8 (27)	0	76 (44)	24 (44)
12 Writing medical records in accordance with legislation	19 (40)	62 (50)	19 (40)	9 (29)	91 (29)	0	37 (50)	53 (51)	10 (32)
13 Writing a letter of discharge	5 (23)	69 (48)	26 (45)	12 (33)	84 (37)	4 (20)	9 (29)	74 (45)	17 (39)
14 Maintaining professional competence	0	38 (50)	62 (50)	21 (41)	58 (50)	21 (41)	17 (39)	31 (47)	52 (51)
15 Building a doctor-patient relationship	5 (22)	67 (48)	28 (46)	14 (36)	81 (40)	5 (22)	25 (44)	70 (47)	5 (22)
16 Working in a team	5 (22)	60 (50)	35 (49)	21 (41)	67 (48)	12 (34)	29 (46)	58 (50)	13 (34)

* also includes paramedics

Table 4

Mean frequency (and standard deviation) of students who reported competences as being addressed during the formal evaluations in percentages per competence

	Internal Medicine*		Surgery			Paediatrics		
Competence	**progress interview**	**final interview**	**progress interview**	**final interview**	**clinical oral**	**progress interview**	**final interview**	**clinical oral**
1 Listing problems and requests for help	46 (51)	62 (50)	39 (50)	25 (44)	69 (48)	64 (49)	61 (50)	74 (45)
2 Taking a history	63 (49)	65 (49)	29 (46)	21 (42)	69 (48)	63 (49)	61 (50)	74 (45)
3 Performing a general physical examination	56 (51)	64 (49)	29 (46)	19 (40)	76 (44)	46 (51)	54 (51)	86 (36)
4 Interpreting and evaluating data	64 (49)	71 (46)	29 (46)	19 (40)	88 (34)	82 (39)	79 (42)	96 (19)
5 Formulating a differential diagnosis	60 (50)	55 (51)	29 (46)	14 (36)	88 (33)	86 (36)	75 (44)	89 (31)
6 Proposing additional investigations	32 (48)	32 (48)	21 (42)	11 (32)	100	46 (51)	61 (50)	89 (31)
7 Outcome of additional investigations	24 (44)	41 (50)	14 (36)	7 (27)	69 (48)	39 (50)	36 (49)	65 (49)
8 Making a management plan	32 (48)	27 (46)	15 (36)	15 (36)	88 (34)	57 (50)	50 (51)	85 (36)
9 Evaluating the result of treatment	17 (38)	24 (44)	4 (20)	0	42 (51)	11 (31)	11 (31)	21 (42)
10 Giving information to the patient	21 (41)	24 (44)	4 (19)	0	13 (35)	12 (33)	12 (33)	33 (48)
11 Writing a letter of referral, or a request for consultation or an additional investigation	16 (37)	23 (43)	7 (27)	4 (20)	13 (34)	7 (26)	7 (26)	19 (40)
12 Writing medical records in accordance with legislation	21 (41)	29 (46)	14 (36)	4 (19)	25 (45)	30 (47)	30 (47)	40 (50)
13 Writing a letter of discharge	8 (28)	5 (22)	7 (27)	4 (19)	0	14 (36)	14 (36)	13 (34)
14 Maintaining professional competence	42 (50)	45 (51)	21 (42)	15 (36)	20 (41)	16 (37)	8 (28)	9 (29)
15 Building a doctor-patient relationship	36 (49)	41 (50)	11 (31)	7 (26)	69 (48)	18 (39)	15 (36)	74 (45)
16 Working in a team	63 (49)	67 (48)	46 (51)	39 (50)	19 (40)	37 (49)	48 (51)	30 (47)

* The clinical oral examination in Internal Medicine was not included in the study

competences, for the progress interview and the final interview, the frequencies were highest in internal medicine and paediatrics and lowest in surgery. The highest frequencies were reported for the clinical oral examination with no differences between surgery and paediatrics. Again, standard deviations were large pointing to strong individual variation among students.

Discussion

We assumed that adequate supervision, feedback and assessment are necessary conditions for effective achievement of clinical competences in the workplace. The results suggest that there is ample room for improvement with regard to each of these factors. For nearly all competences many students reported supervision at a rate that did not reach the minimum benchmark of 35%. This applies particularly for supervision by observation. This confirms findings from a study by Scott *et al*. who reported that most of the time supervision was not based on direct observation and apparently inferred from vicarious information.[21] Another indication that supervision is not a structured educational event is the huge individual variation across students as evidenced by the very large standard deviations. Apparently, supervision in the workplace is a rather haphazard learning event. Because the quality of *feedback* depends on the seniority of the person providing it, we asked students which group provided most of their feedback. Since feedback appeared to be largely provided by residents, serious doubts about the quality of feedback are warranted. In *assessment*, as in supervision, mainly directly patient-related competences were addressed. An exception was team working skills. Interestingly, these skills received a lot of attention in assessment, but remained relatively unattended as far as supervision was concerned. In all, the results appear to support the conclusion that the investigated clerkships do not provide all the necessary conditions for high-quality competence learning in the workplace. The findings in this study bear a strong resemblance to the results of previous studies on procedural and clinical skills.[17,18]

Although the high inter-student variation limits the significance of inter-disciplinary differences, an interesting finding is that students reported relatively more supervision and staff involvement in the internal medicine clerkship. This is in accordance with the fact that internal medicine is the entry clerkship. Perhaps staff of this department are more aware of their responsibility towards undergraduate students. Nevertheless, individual variations between students in their learning experiences are far more impressive than differences between clerkships, which is in accordance with the literature.[22,23]

This study has several limitations. The sample sizes were rather small and the standard deviations very high. Although the latter hamper the interpretation of the results, there are nevertheless important indicators for the quality of learning in the workplace. The fact that we measured the presence of important determinants of the effectiveness of learning in the workplace, meant that we studied the curriculum in action rather than the learned curriculum.[24] In order to substantiate the findings of this study we recommend further research into what students really learn. Our study design did not enable us to elaborate on factors that determine the considerable individual variation in findings across students for supervision, feedback and assessment. A follow-up study in different rotations might clarify whether the differences are mainly determined by student or by rotation. Despite its limitations, this study has provided yet again strong indications that we should not take it for granted that the workplace is an inherently powerful learning environment.

References

1. Societal Needs Working Group, CanMEDS 2000 project. Skills for the new millennium. *Annales CRMCC* 1996;29:206-16.
2. Medical school objectives writing group. Learning objectives for medical students' education—guidelines for medical schools: report I of the medical school objectives project. *Acad Med* 1999;74:13-8
3. Turnbull J, van Barneveld C. Assessment of clinical performance: in-training evaluation. In: Norman GR, van der Vleuten CPM, Newble DI, eds. *International Handbook of Research in Medical Education*. Dordrecht/ Boston/London: Kluwer Academic Publishers 2002;793-810.
4. Metz JCM, Stoelinga GBA, Pels Rijcken-Van Erp Taalman Kip EH, van den Brand-Valkenburg BWM. *Blueprint 1994: Training of doctors in the Netherlands. Objectives of undergraduate medical education in the Netherlands.* Nijmegen: University Publication Office 1994.
5. WFME Task force on defining international standards in basis medical education. Report of the working party. Copenhagen, 14-16 October 1999. *Med Educ* 2000;34:665-75.
6. Karle H. Global standards in medical education - an instrument in quality improvement. *Med Educ* 2002;36:604-5.
7. Simpson JG, Furnace J, Crosby J, Cumming AD, Evans PA, Friedman-Ben-David M, Harden RM, Lloyd D, McKenzie H, McLachlan JC, McPhate GF, Percy-Robb IW, McPherson SG. The Scottish doctor-learning outcomes for the medical undergraduate in Scotland: a foundation for competent and reflective practitioners. *Med Teach* 2002;24:136-43.
8. Cooper H. HIMAA professional competencies project. *Health Inf Manag* 1999-2000;29:130-1.
9. Jolly BC, Macdonald MM. Education for practice: the role of practical experience in undergraduate and general clinical training. *Med Educ* 1989;23:189-95.
10. Irby DM. Teaching and learning in the ambulatory care setting, a thematic review of the literature. *Acad Med* 1995;70:898-931.
11. Remmen R, Denekens J, Scherpbier A, Hermann I, van der Vleuten CPM, van Royen P, Bossaert L. An evaluation study on the didactic quality of clerkships. *Med Educ* 2000;34:460-4.
12. Van der Hem-Stokroos HH, Scherpbier AJJA, van der Vleuten CPM, De Vries H, Haarman HJThM. How effective is a clerkship as a learning environment? *Med Teach* 2001;23:608-13.
13. Kilminster S, Jolly B, van der Vleuten CPM. A framework for effective training for supervisors. *Med Teach* 2002;24:385-9.
14. Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. *Med Educ* 1983;17:165-71.

15. Stillman PL, Haley HL, Regan MB, Philbin MM. Positive effects of a clinical performance assessment program. *Acad Med* 1991;66:481-3.
16. Irby DM. What clinical teachers in medicine need to know. *Acad Med* 1994;69:333-42.
17. Remmen R, Denekens J, Scherpbier AJJA, van der Vleuten CPM, Hermann I, Van Puymbroeck H, Bossaert L. Evaluation of skills training during clerkships using student focus groups. *Med Teach* 1998;20:428-31.
18. Kassebaum DG, Eaglen RH. Shortcomings in the evaluation of students' clinical skills and behaviors in medical school. *Acad Med* 1999;74:842-9.
19. Wolfhagen HAP. *Kwaliteit van klinisch onderwijs* [thesis] Maastricht: Maastricht University 1993.
20. Stern DT, Williams BC, Gill A, Gruppen LD, Woolliscroft JO, Grum CM. Is there a relationship between attending physicians' and residents' teaching skills and students' examination scores? *Acad Med* 2000;75:1144-6.
21. Scott CS, Irby DM, Gilliland BC, Hunt DD. Evaluating clinical skills in an undergraduate medical education curriculum. *Teach Learn Med* 1993; 5:49-53.
22. Gruppen LD, Wisdom K, Anderson DS, Woolliscroft JO. Assessing the consistency and educational benefits of students' clinical experiences during an ambulatory care internal medicine rotation. *Acad Med* 1993;68:674-80.
23. Seabrook MA, Woodfield SJ, Papagrigoriadis S, Rennie JA, Atheron A, Lawson M. Consistency of teaching in parallel surgical firms: an audit of student experiences at one medical school. *Med Educ* 2000;34:292-8.
24. Remmen R. *An evaluation of clinical skills training at the medical school of the university of Antwerp* [thesis] Antwerp: University of Antwerp 1999.

Chapter 6

Effects of an in-training assessment programme on supervision of and feedback on competences in an undergraduate Internal Medicine clerkship

HEM Daelmans, RJI Hoogenboom, AJJA Scherpbier,
CDA Stehouwer, CPM van der Vleuten

Published in *Medical Teacher* 2005 (in press)

Summary

Assessment drives the educational behaviour of students and supervisors. Therefore, an assessment programme targeted at specific competences may be expected to motivate supervisors and students to pay more attention to those competences. In-training assessment (ITA) is regarded as a feasible method for assessing a broad range of competences. Before and after the implementation of an ITA programme in an undergraduate Internal Medicine clerkship we surveyed students on the frequency of unobserved and observed supervision, and the quality of feedback as inferred from the seniority of the person providing it. After the implementation of the ITA programme supervision increased, but the difference was not statistically significant. The quality of feedback showed no significant change either. Inter-student variation in supervision and feedback remained invariably high after the implementation of the ITA programme. Whether these results are attributable to the way the programme was implemented or to the way the results were assessed remains to be clarified.

Overview

The implementation of an ITA programme in an undergraduate Internal Medicine clerkship does not automatically result in a significant increase in the frequency of *supervision* of clinical competences as perceived by students.

Involving teachers of high seniority in the ITA programme does not necessarily produce a significant increase in *feedback* on clinical competences by those teachers as perceived by students.

Individual *variation* across students in perceived supervision and feedback does not necessarily diminish after the implementation of an ITA programme.

Introduction

A competence is the ability of a professional to handle complex situations or problems using professional knowledge, skills and attitudes in an integrative way. Learning in a relevant context is considered to be important for students' acquisition of competences for their future profession.[1-3] The most relevant context for the learning of undergraduate medical students is provided by clinical clerkships. The effectiveness of clerkships as a learning environment depends on variables like supervision, feedback and assessment.[4-9] Key factors in effective acquisition of competences are therefore adequate supervision, feedback and assessment. In a previous study of these key factors in clerkships in three disciplines, we found that supervision of students' clinical competences was scant and mostly unobserved, that most of students' feedback was provided by residents, i.e. trainees receiving their specialised training, and that assessment mainly focused on a limited range of competences.[10] We assumed that senior staff provide feedback of a higher quality compared with residents.[11] The study also revealed considerable variation between individual students in the amount of supervision and feedback received.

Assessment drives the educational behaviour of students and supervisors. It has been described as the most powerful influence on students' learning behaviour.[7,8] Moreover, Stillman reported evidence that, when it is known that a particular competence is likely to be assessed, observation of and feedback on that competence are likely to increase.[9] This increase may be attributable to more requests for observation and feedback by students or to more attention for observation and feedback on the part of the teachers.

In clerkship assessment, it is essential that students' actual performance of clinical competences is assessed.[12,13] However, performance assessment is difficult to achieve and therefore rarely accomplished. Moreover, reliable assessment requires assessment of large samples of competences by a variety of methods and a large group of assessors.[12-17] Those prerequisites constitute a

feasibility problem, which forms a barrier to reliable assessment of clinical competences. Recently, in-training assessment (ITA), defined as systematic observation, feedback and documentation of students' performance during clinical training was proposed as a feasible method to achieve performance assessment.[18] ITA has been advocated as a valid method for the assessment of actual performance in clinical practice. It is feasible for a broad range of competences during clerkships and offers better reliability than do the traditional subjective clinical evaluations.[18,19] It is also possible to involve a large group of assessors in an ITA programme so that the workload per assessor can be reduced.

When ITA is used for a wide range of competences, supervisors may be motivated to supervise students' performance of those competences more frequently. Recruiting a group of staff members as assessors might stimulate senior staff to be more active in giving feedback to students. Students' learning can also be directed towards the required competences by frequent ITAs.

We implemented an ITA programme in an undergraduate clerkship with the aim of increasing the frequency of supervision and the quality of feedback on clinical competences and reducing inter-student differences in these respects. A questionnaire study investigated the effects of the programme.

Method

Context

In January 1999, we implemented an ITA programme in the traditional Internal Medicine clerkship at the Vrije Universiteit Medical Centre (VUmc), Amsterdam, the Netherlands. Students enter the undergraduate medical curriculum of the VUmc immediately after finishing their secondary education. The undergraduate curriculum consists of four preclinical years of interdisciplinary mainly lecture based education followed by two years of discipline-based traditional clerkships. The first clerkship for all students is in Internal Medicine; thereafter the sequence of the clerkships varies.

Participants

Data were collected before and after the implementation of ITA, i.e. between May 1998 and January 1999 and between January 1999 and January 2001, respectively. All students entering the clerkship between May 1998 and January 2001 were asked to participate. Participation was voluntary and entailed completion of a questionnaire in the last week of the clerkship. Students in the traditional clerkship (28) were asked to complete the same questionnaire as the students in the ITA programme (63).

ITA programme

The ITA programme comprised five ITA formats: one student-patient encounter, one critical appraisal session, one case presentation, four structured long cases and twelve case write-ups. Members of staff acted as assessors for the first four formats and residents assessed the case write-ups. Structured rating forms with a five-point Likert scale (1 = fail, 2 = borderline, 3 = pass, 4 = high pass, 5 = excellent) were used to record students' performance. There was space on the forms for narrative feedback.

The internal medicine educational co-ordinator and the first author informed all persons involved in the study about the ITA programme in a letter and during a meeting. Additionally, students were given the structured rating forms for all assessments on the first day of clerkship to familiarise them with the assessment formats. Residents, members of staff and the head of the department received all the structured rating forms by mail and subsequently could ask questions about the ITA programme at a morning report during which the ITA programme was elucidated. In assigning ITA formats to the supervisors, we tried to take account of their preferences and schedules. Some supervisors performed only one ITA format, others performed more depending on the interactions they were able to have with students.

Questionnaire

We developed a questionnaire for this study, asking students to indicate for seven educational events in which supervision was likely to occur how often they had been supervised on each of thirteen competences, whether supervision was unobserved or observed and which category of supervisors provided the most feedback. The educational events were: student-patient contacts, review of medical records, ward rounds, paper rounds, case presentations, critical appraisal sessions and bedside teaching. Non-observed supervision consisted of a discussion of the competence afterwards and or verification of the performance (findings checked at a later time). Supervision was characterised as observed if the supervisor was actually present when the student performed the competence in question. The frequency of supervision was rated on a three-point scale (<35%; at least 35% or <70%; and at least 70%) indicating the percentage of the different educational events on which supervision took place. Thirty-five per cent was an arbitrary minimum benchmark and 70% was considered to be a good score. The quality of feedback was inferred from the seniority of those providing it (Stern et al., 2000), and that is why students were asked to indicate which of three groups of potential feedback providers actually gave most of the feedback: 1) peers, i.e. other students as well as nursing and paramedical staff, 2) residents and 3) academic clinical staff.

The competences included in the questionnaire were derived from 'Blueprint 1994, training of doctors in the Netherlands'.[20] This document is the result of a consensus procedure among Dutch experts in medical education. It contains general competences, discipline-related clinical and procedural skills and problems as starting point for training. A panel of three clinical education co-ordinators selected the competences they considered the most relevant for this study. Of twenty competences thirteen were included in the study (Table 1).

Table 1

Mean frequencies of students (and standard deviation) reporting unobserved and observed supervision in <35% and >70% of the seven selected educational events in percentages per competence before and after the introduction of the ITA programme in the internal medicine clerkship

	Supervision <35% of events				Supervision >70% of events			
	Unobserved		Observed		Unobserved		Observed	
Competence	Before	After	Before	After	Before	After	Before	After
1 Listing problems and requests for help	49 (35)	29 (31)	68 (35)	65 (36)	20 (32)	35 (35)	10 (17)	16 (27)
2 Taking a history	34 (37)	25 (32)	76 (26)	74 (26)	24 (36)	39 (40)	11 (16)	10 (19)
3 Performing a general physical examination	44 (42)	36 (35)	65 (31)	58 (30)	20 (38)	27 (38)	13 (22)	13 (17)
4 Interpreting and evaluating data [a]	23 (34)	30 (33)	-	-	33 (43)	37 (39)	-	-
5 Formulating a differential diagnosis [b]	31 (37)	27 (32)	-	-	27 (35)	37 (36)	-	-
6 Proposing additional investigations	35 (34)	34 (34)	-	-	13 (29)	33 (38)	-	-
7 Outcome of additional investigations [c]	60 (31)	53 (35)	-	-	10 (21)	13 (24)	-	-
8 Making a management plan	45 (41)	35 (38)	-	-	11 (28)	30 (39)	-	-
9 Evaluating the result of treatment	65 (39)	62 (39)	78 (41)	69 (41)	12 (27)	8 (22)	6 (22)	9 (28)
10 Giving information to the patient	-	-	88 (34)	76 (43)	-	-	4 (20)	2 (14)
11 Writing medical records in accordance with legislation	81 (29)	72 (36)	-	-	7 (21)	10 (21)	-	-
12 Maintaining professional competence [d]	51 (44)	54 (42)	33 (48)	34 (48)	14 (27)	17 (29)	33 (48)	40 (49)
13 Building a doctor-patient relationship	86 (27)	61 (39)	73 (32)	59 (41)	0	5 (19)	3 (9)	8 (26)

[a] data from problem description, history and physical examination;

[b] differential diagnosis or problem list, probability diagnosis, or working hypothesis;

[c] interpreting and evaluating the outcome;

[d] advancing and maintaining professional competence through actively searching for relevant literature, reviewing professional literature critically and taking responsibility for one's own lifelong learning

- was not included in questionnaire

Statistical analysis

We calculated the mean frequency of feedback and supervision for each competence across students before and after implementation of the ITA programme. For supervision we calculated the mean frequency of supervision of competences reported by students across the selected educational events as < 35% or 70% or more. For feedback we calculated the mean frequency of students reporting that various supervisor categories were the most involved in the feedback on competences. The significance of the differences between the traditional and the ITA group was tested with the Mann-Whitney test. After a Bonferroni correction for the number of comparisons a p-value of 0.05 was considered significant.

Results

Twenty-five students in the traditional clerkship (response rate 86%) and 51 students in the clerkship with the ITA programme (response rate 81%) returned the completed questionnaire. Non-response was mainly due to organisational shortcomings.

Supervision

In Table 1 mean frequencies and standard deviations are reported for each of the competences for two benchmarks of supervision prior to and after implementation of the ITA programme.

After the implementation of the ITA programme, observed and unobserved supervision of almost all the competences increased, albeit that the increases were not statistically significant. Before and after implementation of the ITA programme supervision was mostly unobserved and concentrated on directly patient-related competences. The standard deviations of students' scores were equally large before and after implementation of the ITA programme.

Table 2

Mean frequency (and standard deviation) of students reporting the supervisor categories providing most feedback in percentages per competence before and after the implementation of the ITA programme in the Internal Medicine clerkship.

Competence	Resident		Staff		Peers*	
	Before	After	Before	After	Before	After
1 Listing problems and requests for help	76 (44)	92 (27)	20 (41)	4 (20)	4 (20)	4 (20)
2 Taking a history	77 (43)	83 (38)	9 (29)	12 (32)	14 (35)	6 (24)
3 Performing a general physical examination	68 (48)	78 (42)	27 (46)	20 (41)	5 (21)	2 (14)
4 Interpreting and evaluating data [a]	70 (47)	90 (30)	30 (47)	8 (27)	0	2 (14)
5 Formulating a differential diagnosis [b]	87 (34)	88 (33)	13 (34)	10 (30)	0	2 (14)
6 Proposing additional investigations	82 (39)	88 (32)	18 (39)	10 (30)	0	2 (14)
7 Outcome of additional investigations [c]	91 (29)	86 (35)	9 (29)	10 (30)	0	4 (20)
8 Making a management plan	74 (45)	92 (27)	26 (45)	6 (23)	0	2 (14)
9 Evaluating the result of treatment	80 (41)	95 (21)	20 (41)	2 (15)	0	2 (15)
10 Giving information to the patient	68 (48)	86 (35)	14 (35)	6 (24)	18 (39)	8 (28)
11 Writing medical records in accordance with legislation	62 (50)	81 (40)	19 (40)	11 (31)	19 (40)	9 (28)
12 Maintaining professional competence [d]	38 (50)	44 (50)	62 (50)	50 (51)	0	6 (24)
13 Building a doctor-patient relationship	67 (48)	82 (39)	28 (46)	9 (29)	5 (22)	9 (29)

* also includes paramedic staff

[a] data from problem description, history and physical examination;

[b] differential diagnosis or problem list, probability diagnosis, or working hypothesis;

[c] interpreting and evaluating the outcome;

[d] advancing and maintaining professional competence through actively searching for relevant literature, reviewing professional literature

Feedback

Table 2 shows mean frequencies and standard deviations of supervisor categories involved in feedback.

After implementation of the ITA programme the percentage of students reporting peers or residents as their main feedback providers increased, with a concomitant decrease in the percentage of students reporting academic staff as the main feedback providers. The differences were not significant, however. The standard deviations remained as high as before the implementation of the ITA programme.

Discussion

For a set of selected clinical competences, we investigated whether the implementation of an ITA programme in an Internal Medicine clerkship improved the frequency of supervision and the quality of feedback received by students and diminished inter-student variation in feedback and supervision.

The design of the study reported in this paper has some weaknesses, which may have impacted on the results. First, we used a pre-post study design. We studied two different groups of students during two different periods. Despite the uniformity of the group of supervisors and the absence of other educational changes besides the ITA programme, we cannot rule out the possibility that the results do not exclusively reflect the impact of the ITA programme. Second, the pre-intervention group was quite small. Given the high standard deviations, differences in outcomes between the two study periods would have to be substantial to reach statistical significance. It is therefore not surprising that the differences that emerged were not significant. This is a problem that needs to be addressed in further research. Third, we investigated students' perceptions of the amount of supervision and feedback they received from supervisors. We did not record actual events. Some events may not have been recognized as supervision or feedback by the students, which may have resulted in

underreporting of the actual supervision and feedback. Fourth, we measured the effectiveness of workplace learning by determining whether important requirements for effectiveness were being met. In doing so we studied the curriculum in action (the teaching offered to the students), but we did not study the learned curriculum (the educational effects of the curriculum on the students).[21] Given these drawbacks no firm conclusions can be drawn on the basis of the interpretations of the results that are discussed below. However, the results offer important leads for further research on this topic.

After the implementation of the ITA programme, we found no statistically significant increase in the frequency of *supervision* of clinical competences nor an increase in feedback by persons with high clinical seniority or a decrease in inter-student variation in these respects. Our expectations of the ITA programme, i.e. improvements in the frequency of supervision and quality of feedback and reduction of inter-student variation on these points, were not borne out by the results.

There may have been flaws in the implementation of the ITA programme that prevented any significant effects on supervision. Supervisors were only informed about the programme at the start in January 1999. Some supervisors assessed only one or two ITA formats. Despite our expectation that all supervisors would be more attentive to the competences included in the ITA programme, it was actually quite easy for supervisors to forget about competences other than those for which they had signed up as an assessor. Perhaps supervisors' awareness of the programme could have been reinforced by bringing the programme to their attention repeatedly, for instance at morning or evening reports through discussion of problems encountered and assessment outcomes. Heightened awareness might have been reflected in more frequent supervision.[22] We provided the students with structured rating forms for all assessments to alert them to the competences included in the programme thereby encouraging them to actively seek supervision on those competences. The results offer no proof either way about the relationship between awareness and active supervision seeking. However, if students did indeed seek more

supervision, their attempts were not successful in terms of significant changes in the reported frequency of supervision. In addition, the supervision provided during the ITAs did not have a significant effect on the reported frequency of supervision.

We expected that the programme would stimulate staff to give students more feedback on competences during everyday clinical work. Moreover, the ITA formats had inbuilt opportunities for more student staff contacts, which we hoped might facilitate requests for feedback by students both during the ITA's and during everyday clinical work. It is obvious from the results that this is not what happened. Members of staff are known to have only limited time for teaching and giving feedback to students.[23,24] Perhaps the extra educational role of assessor in the ITA programme may even have had an adverse effect by detracting from the available teaching time and thus diminishing feedback opportunities.

The ITA programme had no demonstrable effect on inter-student variation in supervision and feedback. As before the programme, differences between students remained invariably high. Apparently, the influence of individual differences between students on students' clinical experiences outweighed any impact of the ITA programme. Other studies have also shown that such differences are very hard to change.[25] Despite unequivocal proof that assessment drives learning behaviour, the effects of particular tests on student learning are not entirely predictable.[7,8,26,27] The high individual variation limited the interpretation of differences before and after the ITA programme was implemented.

From the literature on continuing medical education we know that a single intervention is unlikely to be successful in terms of behavioural change and interventions should target as many factors as possible.[28] However, considering that assessment is potentially a very powerful drive for students' and possibly also teachers' educational behaviour, we considered it defensible to limit this study to a single intervention.[7-9] Further research will be necessary to explain why the ITA programme had no demonstrable significant effects on supervision and feedback. Such a study might be quantitative with larger sample sizes. It would also be preferable to

limit the study to one period of time and divide the students and the assessors into groups with ITA and groups without ITA. However, in practice this may not be feasible.[29] Therefore a qualitative study focusing on the effects of the ITA programme on teachers' and students' behaviour might be a good alternative.

Conclusion

The implementation of an ITA programme in an undergraduate Internal Medicine clerkship did not yield the expected improvement in the frequency of supervision and the quality of feedback nor the expected reduction in individual differences between students in those respects. Further research will be necessary to shed more light on the effects of the ITA programme on students' and supervisors' behaviour.

Acknowledgements

The authors wish to thank Prof. A.J.M. Donker, MD, PhD, emeritus professor of Internal Medicine, for enabling the implementation of the ITA programme in the Department of Internal Medicine. The authors also want to express their gratitude to A. Thijs, educational co-ordinator of Internal Medicine, for supporting the implementation of the programme and the questionnaire survey.

References

1. American Health Information management Association Position statement. Issue: Educational clinical affiliations. *Journal of AHIMA* 1995;66:following S104.
2. Headrick LA, Neuhauser D, Schwab P, Stevens DP. Continuous quality improvement and the education of the generalist physician. *Acad Med* 1995;70:S104-9.
3. Regehr G, Norman GR. Issues in cognitive psychology: implications for professional education. *Acad Med* 1996;71:988-1001.
4. Irby DM. What clinical teachers in medicine need to know. *Acad Med* 1994;69:333-42.
5. Jolly BC, Macdonald MM. Education for practice: the role of practical experience in undergraduate and general clinical training. *Med Educ* 1989;23:189-95.
6. Kilminster S, Jolly B, van der Vleuten CPM. A framework for effective training for supervisors. *Med Teach* 2002;24:385-9.
7. Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. *Med Educ* 1983;17:165-71.
8. Messick S. The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* 1994;23:13-23.
9. Stillman PL, Haley HL, Regan MB, Philbin MM. Positive effects of a clinical performance assessment program. *Acad Med* 1991;66:481-3.
10. Daelmans HEM, Hoogenboom RJI, Donker AJM, Scherpbier AJJA, Stehouwer CDA, van der Vleuten CPM. Effectiveness of clinical rotations as a learning environment for achieving competences. *Med Teach* 2004; 26:305-12.
11. Stern DT, Williams BC, Gill A, Gruppen LD, Woolliscroft JO, Grum CM. Is there a relationship between attending physicians' and residents' teaching skills and students' examination scores? *Acad Med* 2000;75:1144-6.
12. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990;65:S63-7.
13. Wass V, van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. *The Lancet* 2001;357:945-9.
14. Kassebaum DG, Eaglen RH. Shortcomings in the evaluation of students' clinical skills and behavior in medical school. *Acad Med* 1999;74:842-9.
15. Schuwirth LWT, Southgate L, Page GG, Paget NS, Lescop JMJ, Lew SR, Wade WB, Baron-Maldonado M. When enough is enough: a conceptual basis for fair and defensible practice performance assessment. *Med Educ 2002;*36:925-30.
16. Van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ Theory Pract* 1996;1:41-67.

17. Swanson DB. A measurement framework for performance-based tests. In: Hart IR, Harden RM, eds. *Further Developments in Assessing Clinical Competence.* Montreal: Can Heal Publications 1987;13-45.
18. Turnbull J, MacFadyen J, van Barneveld C, Norman G. Clinical work sampling: a new approach to the problem of in training evaluation. *J Gen Int Med* 2000;15:556-61.
19. Hays R, Wellard R. ITA in postgraduate training for general practice, *Med Educ 1989;*32:507-13.
20. Metz JCM, Stoelinga GBA, Pels Rijcken-van Erp Taalman Kip EH, van den Brand-Valkenburg BWM. *Blueprint 1994; Training of doctors in the Netherlands. Objectives of undergraduate medical education in the Netherlands*. Nijmegen: University Publication Office 1994.
21. Remmen, R. *An Evaluation of Clinical Skills Training at the Medical School of the University of Antwerp* [thesis] Antwerp, University of Antwerp 1999.
22. Spike N, Alexander H, Elliot S, Hazlett C, Kilminster S, Prideaux D, Roberts T. In-training assessment - its potential in enhancing clinical teaching. *Med Educ* 2000;34:858-61.
23. Samuel S, Shaffer K. Profile of medical student teaching in radiology: teaching methods, staff participation, and rewards. *Acad Radiol 2000;* 7:868-74.
24. Regan-Smith M, Young WW, Keller AM. An efficient and effective teaching model for ambulatory education. *Acad Med* 2002;77:593-9.
25. Van der Hem-Stokroos HH, Scherpbier AJJA, van der Vleuten CPM, de Vries H, Haarman HJThM. How effective is a clerkship as a learning environment? *Med Teach* 2001;23:608-13.
26. Van Luijk SJ, van der Vleuten CPM, Schelven RM. The relation between content and psychometric characteristics in performance-based testing. In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP, eds. *Teaching and Assessing Clinical Competence.* Groningen: BoekWerk Publications 1990;202-7.
27. Van der Vleuten CPM, Scherpbier AJJA, Dolmans DHJM, Schuwirth LWT, Verwijnen GM, Wolfhagen HAP. Clerkship assessment assessed. *Med Teach* 2000;22:592-600
28. Mann KV. Continuing medical education. In: Norman GR, van der Vleuten CPM, Newble DI, eds. *International Handbook of Research in Medical Education.* Dordrecht/Boston/London: Kluwer Academic Publishers 2002;793-81.
29. Norman G. Editorial- the effectiveness and the effects of effect sizes. *Adv Health Sci Educ Theory Pract* 2003;8:183-7.

Chapter 7

In training assessment: effects on supervision and feedback, a qualitative study

HEM Daelmans, RM Overmeer, HH van der Hem-Stokroos, AJJA Scherpbier, CDA Stehouwer, CPM van der Vleuten

Reviewed by *Medical Education*

Summary

Supervision and feedback are essential factors for the learning environment in workplace learning and their frequency and quality can be improved. Assessment is a powerful tool to influence students' learning and supervisors' teaching and thus the learning environment.
We investigated an in-training assessment programme 'in action' and explored its effects on supervision and feedback using individual semi-structured interviews.
We interviewed eight students and seventeen assessors (nine members of staff and eight residents) in the Internal Medicine undergraduate clerkship at the Vrije Universiteit Medical Centre, Amsterdam, the Netherlands.
The assessment programme 'in action' differed from the 'intended' programme. Assessors provided hardly any follow-up on supervision and feedback given during assessments. Although students wanted more supervision and feedback, they hardly ever asked for it. Students and assessors failed to integrate the whole range of competences included in the in-training assessment programme into their respective learning and supervision and feedback. When giving feedback, assessors rarely gave borderline or fail judgements.
For the intended assessment programme to be congruent with the assessment programme 'in action' the implementation of an ITA programme needs to be monitored and full and repeated information about the programme is necessary. Introducing an ITA programme that includes assessment of several competences does not automatically lead to more attention for these competences in supervision and feedback. Measures that facilitate change in the learning environment seem a prerequisite for enabling the assessment programme to steer the learning environment.

Overview

What is already known
The learning environment in undergraduate clinical rotations:
· is not always optimal.
· can be improved when an assessment programme steers the learning of students and the teaching of supervisors into the desired direction.

What this study adds
In a clinical rotation:
· participants must be adequately and repeatedly informed about a new assessment programme to prevent substantive differences between the 'intended' programme and the programme 'in action'.
· measures to facilitate change in the learning environment are needed to achieve a positive effect of an assessment programme on supervision and feedback.
· assessors are reluctant to express borderline or fail judgements.

Suggestions for further research
Studies should address measures for:
· facilitating changes in the learning environment in clinical rotations.
· changing the negative connotation of borderline and fail judgements.

Introduction

In-training assessment (ITA), defined as systematic observation, feedback and documentation of students' performance during clinical training, has been advocated as a valid assessment method in undergraduate and postgraduate training.[1,2] ITA offers better reliability than traditional subjective clinical evaluations and has proven feasibility for the assessment of a broad range of clinical competences.[3-5] To achieve acceptable reliability, many different assessments by different assessors are required.[6] To achieve feasibility, ITA should be incorporated into the day-to-day work routine of both students and assessors and not take up too much time. A reliable and feasible ITA programme thus incorporates frequent assessments (moments of supervision and feedback) of a broad range of competences by many assessors during day-to-day work.

Supervision and feedback appear to be essential factors in workplace learning.[7,8] Workplace learning is the principal form of learning in undergraduate and graduate clinical rotations. Studies have indicated that the effectiveness of workplace learning is not always optimal and that supervision and feedback occur rather infrequently during clinical clerkships.[9-13] Incorporating an ITA programme, i.e. many structured moments of supervision and feedback, into a clerkship might improve the effectiveness of clerkship learning.

Assessment is often described as the most powerful drive of student learning behaviour.[14,15] Moreover, the knowledge that a certain competence will be assessed may increase supervision and feedback regarding that competence.[16] Since the competences included in an ITA programme are known in advance, such a programme may stimulate students and supervisors to direct their efforts at those competences.[14,16] In this way, the ITA programme may influence students' learning as well as content and frequency of assessors' supervision and feedback.

In addition to frequency and content, the quality of supervision and feedback is important. The quality of supervision may benefit from clearly defined objectives, from structure and from continuity of assessment. Reflection on performance and, for beginners, direction on performance are also considered important.[7] An ITA programme can provide occasions for structured supervision and feedback on the basis of clearly defined objectives. Structured moments for verbal feedback create excellent opportunities for reflection and directive comments on students' performance. In this way, an ITA programme may well enhance the quality of supervision and feedback.

We investigated whether ITA can actually improve the effectiveness of the learning environment. To this end we explored how an ITA programme in an undergraduate medical rotation affected supervision and feedback from the perspectives of both supervisors and students. The ITA programme we studied has been described as feasible and fairly reliable.[5] However, since the 'intended curriculum' is not necessarily congruent with the 'curriculum-in-action', we

examined the actual delivery of the ITA programme before studying its effects on supervision and feedback.[17]

The principal focus of this study was students' and supervisors' actual behaviour and background information about that behaviour. We used a qualitative study design involving semi-structured interviews with individual students and assessors.[18,19] Our main research questions were: 1) What happens in the day-to-day reality of the ITA programme (i.e. in the curriculum in action)? 2) What are the main characteristics of supervision and feedback in the clinical rotation where the ITA programme is operational?

Method

Context

In January 1999, an ITA programme was introduced in the internal medicine clerkship at the Vrije Universiteit Medical Centre (VUmc), Amsterdam, the Netherlands. The first four years of the curriculum consist of interdisciplinary, mainly lecture-based preclinical education, which are followed by two years of discipline-based, traditional clerkships. The first clinical rotation for all students is in internal medicine. The students spend six weeks on the wards where they are supervised by residents and four weeks in the outpatient clinic where they are supervised by residents and staff. Both residents and staff are involved in ITA activities on the wards and in the outpatient clinic.

ITA programme

The ITA programme comprises nineteen assessments spread over five different formats and covers a broad range of competencies (Table 1). Student performance is rated on a five-point Likert scale (1=fail, 2=borderline, 3=pass, 4=high pass, 5=excellent) on structured rating forms with space for narrative feedback. Assessors are explicitly invited to provide written feedback on items with borderline or fail ratings. Before implementation of the programme

Table 1
Numbers, formats, assessors and contents of the different components of the ITA-programme

Component of the ITA-programme	Number of assessments	Assessment format	Assessor	Competencies assessed
Student-patient encounter	1	Direct structured observation, rating and feedback on ward	Member of staff	History taking, physical examination, diagnostic reasoning, verbal communication
Critical appraisal session	1	Direct structured observation, rating and feedback on ward	Member of staff	Scientific relevance of the presented information, quality of critical thinking, verbal communication, presentation
Case presentation	1	Direct structured observation, rating and feedback in out-patient clinic	Member of staff	Completeness and relevance of the presented information, quality of critical thinking, verbal communication, presentation
Case write-up	12	Structured discussion and rating of write-ups on ward and out-patient clinic	Resident	Diagnostic reasoning, management, written communication
Structured long case	4	Structured discussion and rating after presentation in out-patient clinic	Member of staff	History taking, physical examination, diagnostic reasoning, management, written communication, critical appraisal

the assessors were informed about it in writing. Furthermore, they were invited to a meeting where the ITA-programme was discussed. After implementation, assessors could put any queries to the clinical co-ordinator. Assessors receive no special training for the ITA-programme. Students are verbally informed about the ITA programme on the first day of their rotation and receive a written summary of this information and the structured rating forms. They are instructed to give the relevant form to the assessor for each assessment. The clinical co-ordinator contacts students regularly and is available for questions about the programme.

Subjects and Procedure

We interviewed three groups involved in the ITA programme: students, staff and residents. Nine staff members (75%) were selected from those who supervised students regularly. They were selected to represent differences in seniority, functions (clinical director, education co-ordinator) and work settings (wards/ outpatient clinic). To balance group sizes, we also selected nine residents (69%) and nine students (31% of the clerks in the department in that year). The residents were selected for variation in seniority and work settings and the students were randomly selected from the clerks in the department at the time of the study. All accepted an e-mail invitation to participate, which included information about the study. At the interview informed consent was obtained verbally. The students were interviewed in the last week of the rotation, i.e. after having completed the ITA programme. One resident was not interviewed because of logistic problems, which left a study group of eight residents.

The interviews lasted 45-90 minutes and were conducted in the summer of 2003. Time and location were planned for the interviewee's convenience and interruptions were kept to a minimum. All participants consented to tape-recording of the interview.

Interviews

We used semi-structured interviews, because they allow open-ended questions and individualised probing, thus enabling exploration of specific topics without rigidly confining the interview to those topics.[19,20]

In order to focus the exploration of the research questions, we developed an interview guide that consisted of the following open-ended questions: 1. What are your tasks in the ITA programme? 2. Describe the supervision and feedback you received (students) or provided (assessors).

Analysis

Verbatim transcriptions of the interviews were made and anonymised. The first author read three transcripts of each group and designed a coding system using the emerging themes. Next, the first two authors coded three other transcripts of each group, using the themes. The coding was discussed and some minor changes were made. The resulting coding system was identical for the interviews with staff and residents, whereas for the student interviews more themes were added, mainly pertaining to the way students actually achieved competences. The first two authors then read and coded all transcriptions. After some discussion consensus was reached about the coding of all interviews.

Results

The themes of the coding system were joined and categorised according to the two research questions. For every theme we successively present the results for students, staff and residents.

What happens in the day-to-day reality of the ITA programme (curriculum in action)?

Students were informed about the ITA programme on their first day of clerkship. They reported marginal quality of information about the actual assessments, the preparation expected of them and consequences of the ITA. One student said:
"Well, I was informed very briefly. I was handed the register, and I was told there were forms in it and I had to hand it in again at the end of the clerkship. Things were not explained clearly, not at all!" (student 7)

Staff mentioned a similar lack of information. Particularly those who had joined the department after January 1999 had not been informed about the ITA programme. Nevertheless, all members of staff had familiarised themselves with the rating forms.

Residents did not recall receiving any official information at the start of the ITA programme. Students informed them that, in addition to discussing the medical record, they had to discuss and rate case write-ups. Although an additional instructional plenary session for residents was held by the educational co-ordinator, the residents were only familiar with the rating forms for the case write-ups and not with the other ones. One resident said:
"I have absolutely no idea what they have to do during the other assessments by members of staff"(resident 1)

Despite the scant information, all nineteen assessments of the ITA programme were performed. The *students* paid hardly any attention to the rating forms before the assessments. The assessors used the rating forms during assessments, but often failed to discuss all items with students. *Staff* mainly focused on diagnostic and clinical reasoning skills and wanted knowledge items about these skills to be added to most rating forms. The *residents* mainly focused on diagnostic reasoning skills; they advocated inclusion of items about verbal reporting of relevant information in the rating form for the case write-ups.

What are the main characteristics of supervision and feedback in this clinical rotation?

Although *students* were mostly supervised by residents in their day-to-day work, they said that residents were very busy with their own work and spent too little time giving supervision and feedback**.** In the outpatient clinic, supervision occurred more regularly than on the ward. In both settings, supervision by residents focused primarily on direct patient care and much less on student performance. One student said:
"The resident was busy surviving himself and I thought yo, what about me?" (student 4).
Except for specific occasions, like assessments and bedside teaching, students almost never worked with staff.

For the *students,* the focus was on knowledge and patient-related skills. They tried to fill gaps in their knowledge by consulting books and the Internet. They enjoyed improving their patient-related skills by discussing findings from history and physical examination with supervisors and during bedside teaching. However, they were reluctant to reveal gaps in knowledge and skills to residents and staff and rarely asked for supervision, except when they performed a technical skill for the first time or had doubts about findings from history and physical examination. Moreover, on feedback occasions, they were afraid to ask for more feedback than they received. They gave several reasons for this:
"I'm afraid they'll think I'm stupid" (student 9), "I think, well, I'm old and wise enough to look things up for myself" (student 1) and "After all, I've been on this rotation for ten weeks now..." (student 4).

Most students felt hesitant to ask staff for feedback and some students even had difficulty asking their resident for feedback. This caused considerable uncertainty among students about what supervisors actually thought of their performance.

Although they hardly ever asked for it, students did want more supervision and feedback. They expected substantial benefits from more (short) observations, because the supervisor then could

"exactly check what I'm doing" (student 6); "give me guidance on what to do and reflect with me" (student 3); "immediately correct me when I do things wrong" (student 7).

Members of staff supervised students during assessments (one to two per student) and bedside teaching (two per group of students). The paucity of supervisory interactions with students made staff feel unable to provide substantial feedback to students or follow-up on students' reception of feedback. Staff focused primarily on knowledge. Despite agreeing that supervision and feedback were important for a range of competences, few members of staff actually managed to supervise students on these competences outside the scheduled assessments.

One member of staff said:

"The ITA programme helps to give more explicit attention to the student. I observe students a little bit more ..I find it difficult to arrange moments to observe student performance and I often feel guilty when I do not manage to do so". (staff 1)

Residents mainly focused on diagnostic reasoning skills, such as questions to ask during history taking and facts to present in verbal presentations of patient cases. Observation occurred rarely, except when students performed a technical skill for the first time. Findings in newly admitted patients were checked but students' skills hardly observed. Although residents would be able to involve students in their activities and follow-up on feedback, they rarely did so. Statements like "I sometimes make suggestions, but it's not my task to tell them to do things my way" (resident 7) and "when they stay with me as much as possible, they will see what I do and learn from it" (resident 6) indicate that residents were often unaware of the importance of their supervisory role.

Both *residents and staff* gave special attention when they noticed major gaps in a student's competence. Residents supervised these students more often and staff members gave more elaborate feedback during assessments. Borderline or fail judgements were rare, however. The residents felt that their personal relationship with students impeded negative judgements as did their feeling that sometimes they were barely more proficient than the students.

Members of staff faced the problem that the non-existence of a remedial programme meant that they were responsible for arranging an extra rotation for a student they failed. Two comments illustrate these dilemmas.
"Sometimes I want to give a borderline rating but then I don't, because I find a poor result not motivating for the student" (resident 7).
"I admit that I tend to give these students a pass. Unless they do life-threatening things. And in those cases I hope those students will not want to become an internist or general practitioner" (staff 8).

Discussion

This qualitative study of the effects of an ITA programme on supervision and feedback focused on the ITA programme 'in action' and on the main characteristics of supervision and feedback

All assessments were performed. However, the results suggest that the paucity of information to students and assessors about the ITA programme adversely affected students' and assessors' preparation and performance in the assessments. This seems to support the desirability of measures to ensure full and repeated information to all participants about the ITA programme to avoid discrepancies between the 'intended' programme and the programme 'in action'. Monitoring the programme's implementation seems equally important.

The four principal findings concerning supervision and feedback are: 1. supervision and feedback were sparse; 2. students hardly ever asked for supervision and feedback; 3. students and assessors only rarely managed to integrate the ITA competences into learning, supervision and feedback; and 4. assessors were reluctant to give true but painful feedback when students' performance merited a rating as borderline or fail.

Supervision and feedback were largely confined to moments when they were unavoidable. Lack of time appeared to hamper more frequent supervision and feedback. Follow-up on feedback was impeded by the scarcity of student staff interactions. Staff members might schedule meetings with students to follow up on earlier

feedback in the outpatient clinic where they supervise residents and students regularly.[21] Despite their more frequent student contacts, residents almost never followed up on feedback due to lack of time and failure to appreciate the importance of their supervisory role. Information about this role might encourage residents to intensify supervision and feedback.

Students rarely asked for supervision and feedback. Earlier studies also revealed reluctance among students to request feedback.[9,22] Apparently, students did not feel safe enough to expose their uncertainties to clinical teachers. The lack of follow-up feedback on students performance due to time constraints may inhibit requests for supervision and feedback. Students will only reveal uncertainties when they know that any future progress will be noticed. More follow-up on feedback may contribute to a learning environment in which students feel free to invite feedback.

The ITA programme was not integrated in the rotation. Students kept focussing on knowledge and patient-related skills, such as diagnostic and therapeutic reasoning skills. Residents and staff continued to focus on knowledge and diagnostic and clinical reasoning skills. This suggests that the introduction of an ITA programme addressing a fairly broad range of competences does not automatically ensure that assessors and students will address those competences. We did not explore in detail why students and assessors did not change their focus of attention. Nevertheless, adequate and repeated information to all parties involved about assessment content and consequences seems a prerequisite for effecting changes in the direction of student learning and supervisor teaching.[14,16] Information might be provided during or even before implementation of the programme, for instance by consulting a group of students and supervisors on the design of the rating forms. Once the ITA programme is in place, information must continue to be provided as new assessors and students enter the rotation.

The ITA programme failed to promote feedback about inadequate performance. Supervisors tended to positively distort student ratings, because students would see the ratings and low ratings were considered demotivating. In their review, Williams *et al*.

observed this mechanism in research on social psychology and business management.[23] Lack of occasions to follow-up on feedback and discuss matters with students may be to blame. This undesirable situation might be improved by measures to alter the negative connotations of borderline and fail judgements by changing the perspective in clerkship learning from 'failing and passing performance' to 'growth in performance'. Although difficult to organise, continuity of supervision and feedback are prerequisites for a changed perspective on judgements. Rewarding honest negative judgements and helping assessors in developing a remedial programme for students might help, as might teacher training in giving borderline or fail judgements.

Conclusion

The introduction of an ITA programme that includes assessment of several competences does not automatically boost supervision and feedback on these competences. Whether an ITA programme will enhance the frequency, content and quality of supervision and feedback will depend on measures that promote information and continuity of supervision and feedback. Early, adequate and repeated information to all stakeholders about content and consequences of the ITA programme and involvement of stakeholders in constructing the programme would enable consultation of supervisors and enlisting their support. Continuity of supervision and feedback might encourage supervisors to provide clear and honest feedback and follow up on that and it might encourage students to ask for supervision and feedback, because they know that improvement of inadequate competences is possible.

In summary, the results of our study suggest that successful implementation of an ITA programme in an undergraduate rotation requires careful monitoring and measures to facilitate changes in the learning environment. More research on such measures and their effects is recommended.

Acknowledgements

The authors wish to thank the members of staff, residents and the students who agreed to be interviewed. Acknowledgements are also due to Mereke Gorsira for linguistic support.

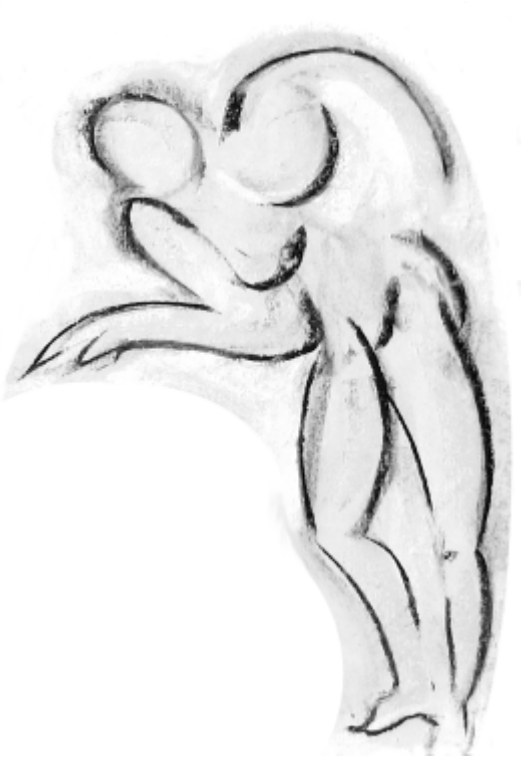
References

1. Turnbull J, Van Barneveld C. Assessment of clinical performance: in-training evaluation. In: Norman GR, van der Vleuten CPM, Newble DI, eds. *International Handbook of Research in Medical Education.* Dordrecht/ Boston/London: Kluwer Academic Publishers 2002; 793-810.
2. Feletti G, Cameron D, Dawson-Saunders B, Des Groseilliers JP, Dooley B, Farmer E, McAvoy P. In-training assessment. In: Newble D, Jolly B, Wakeford R, eds. *The Certification and Recertification of Doctors: Issues in the Assessment of Clinical Competence.* Cambridge: Cambridge University Press 1994;151-66.
3. Hays R, Wellard R. ITA in postgraduate training for general practice. *Med Educ* 1998;32:507-13.
4. Turnbull J, MacFadyen J, Van Barneveld C, Norman G. Clinical work sampling: a new approach to the problem of in training evaluation. *J Gen Intern Med* 2000;15:556-61.
5. Daelmans HEM, Van der Hem-Stokroos HH, Hoogenboom R, Scherpbier AJJA, Stehouwer CDA, Van der Vleuten CPM. Feasibility and reliability of an in-training evaluation programme in an undergraduate clerkship. *Med Educ* 2004;38:1270-7.
6. Norman GR. Objective measurement of clinical performance. *Med Educ* 1985;19:43-7.
7. Kilminster S, Jolly B, Van der Vleuten CPM. A framework for effective training for supervisors. *Med Teach* 2002;24:385-9.
8. Irby DM. What clinical teachers in medicine need to know. *Acad Med* 1994; 69:333-42.
9. Irby DM. Teaching and learning in ambulatory care settings: a thematic review of the literature. *Acad Med* 1995;70:898-931.
10. Jolly BC, Macdonald MM. Education for practice: the role of practical experience in undergraduate and general clinical training. *Med Educ* 1989;23:189-95.
11. Remmen R, Denekens J, Scherpbier A, Hermann I, Van der Vleuten CPM, Van Royen P, Bossaert L. An evaluation study on the didactic quality of clerkships. *Med Educ* 2000;34:460-4.
12. Van der Hem- Stokroos HH, Scherpbier AJJA, Van der Vleuten CPM, De Vries H, Haarman HJThM. How effective is a clerkship as a learning environment? *Med Teach* 2001;23:608-13.
13. Remmen R, Denekens J, Scherpbier AJJA, Van der Vleuten CPM, Hermann I, Van Puymbroeck H, Bossaert L. Evaluation of skills training during clerkships using student focus groups. *Med Teach* 1998;20:428-31.
14. Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. *Med Educ* 1983;17:165-71.

15. Messick S. The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* 1994;23:13-23.
16. Stillman PL, Haley HL, Regan MB, Philbin MM. Positive effects of a clinical performance assessment program. *Acad Med* 1991;66:481-3.
17. Remmen R. *An Evaluation of Clinical Skills Training at the Medical School of the University of Antwerp* [thesis]. Antwerp: University of Antwerp 1999.
18. Patton MQ. *Qualitative Evaluation and Research Methods*. London: Sage 1990.
19. Pope C, Mays N. Qualitative research: reaching the parts other methods cannot reach: an introduction to qualitative methods in health and health services research. *BMJ* 1995;311:42-5.
20. Britten N. Qualitative research: qualitative interviews in medical research. *BMJ* 1995;311:251-3.
21. Regan-Smith M, Young WW, Keller AM. An efficient and effective teaching model for ambulatory education. *Acad Med* 2002;77:593-9.
22. Van der Hem-Stokroos HH, Daelmans HEM, Van der Vleuten CPM, Haarman HJThM, Scherpbier AJJA. A qualitative study of constructive clinical learning experiences. *Med Teach* 2003;25:120-6.
23. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teach Learn Med* 2003;15:270-92.

Chapter 8

Discussion



Introduction

Competence learning is becoming increasingly important in undergraduate and graduate education. Competence learning can be integrated in the training programme of the early years of undergraduate medical education, for instance by offering as much relevant context as possible in tasks students have to perform and by taking care that these tasks require students to integrate knowledge, skills and behaviour. The workplace setting of the clinical clerkships offers a far more authentic learning environment, however, and is thus indispensible for putting the finishing touches to competence learning. Workplace learning in clinical, community or out-patient care has some weaknesses with respect to the learning environment.[1-4] Assessment has been described as an important influence on students' and teachers' behaviour and it can be used as a tool to improve the learning environment.[5-7] There is a need for new methods for adequate assessment of performance in the workplace.[8-10] Recently, in-training assessment was introduced as a new method to assess performance. In-training assessment has been defined as multiple, structured and observed assessments, documented immediately following the assessment and performed throughout a clinical rotation.[11-13] Although this assessment method has a lot of potential for assessing a broad range of competences, research is needed to identify formats that offer improved reliability and acceptable feasibility, which would enable broad use of this assessment method.

In this thesis two research themes were addressed. Firstly, the effects of an in-training assessment programme on the learning environment in an undergraduate clerkship and, secondly, feasibility and reliability aspects of this in-training assessment programme. Because the assessment programme needed to be established before its effects on the learning environment could be studied, the feasibility and reliability of the in-training assessment programme will be dealt with first. For each theme, findings and discussion on the detailed research questions are presented and methodological considerations are given. Subsequently, general findings are

discussed for each theme and recommendations for further research are presented. At the end of the discussion, the general conclusions and implications of this thesis are presented.

Feasibility and reliability aspects of in-training assessment

Findings and discussion

Research question 1: What is the feasibility and reliability of multiple long case clinical examinations in an undergraduate clinical clerkship? (chapter 2)

Multiple long case clinical examinations were implemented in an undergraduate clinical clerkship. The examinations were scored in two ways, i.e. on a detailed item scale and as a single global judgement. It proved possible to construct a multiple long case clinical examination that was both feasible and reliable. Although a sizeable time investment was needed to reach sufficient reliability, this investment appeared not to differ from that needed for other test formats, such as OSCEs.[9,14] The reliability of the global judgements was higher than that of the averaged item grades, indicating that the global judgements were including other aspects of student functioning than the item scores. Global judgements thus seem to add information to assessment results that are based on item ratings. This finding seems to confirm the results of earlier research, which pointed in the same direction.[15,16]

Multiple long case clinical examinations could reach sufficient reliability and feasibility, especially when global judgements are used. However, considering the arguable validity of this type of assessment, it would be ill advised to focus exclusively on long cases in the assessment of clinical performance.[17] Long case clinical examinations involve an unobserved student patient encounter. Observation of parts of the history and/or physical examination could improve the validity of the long case clinical

examination.[18] Moreover, combining long case clinical examinations with a mix of other test formats would enable assessment of a broad range of competences.[10,20-22] It seems therefore advisable to combine the multiple long case clinical examination with a range of other test formats, including formats that involve observation of the competences that are assessed, either within or in addition to the long case.

Research question 2: What is the feasibility and reliability of the in-training assessment programme for assessing students' clinical performance? (chapter 3)

An in-training assessment programme consisting of five test formats, i.e. three single-sample formats (student-patient encounter, critical appraisal session and case presentation) and two multiple-sample formats (twelve case write-ups and four structured long cases) was implemented in an undergraduate clinical clerkship. In all, the in-training assessment programme included nineteen assessments. The in-training assessment programme proved to be both feasible and reliable. Reliability for nineteen assessments was high (0.81) but probably overestimated, because the outcome was strongly affected by the multiple-sample formats. When reliability was estimated for five test formats, the result was somewhat disappointing (0.55). This was most probably due to low reliability of the single-sample formats. Reliability would probably benefit from replacing single-sample test formats with multiple-sample formats.[9,23] We used single-sample formats of tests that were thought to take up a lot of examiners' time. This means that feasibility might have been seriously hampered if we had included these tests as multiple-sample formats. An option to increase reliability without jeopardising feasibility might have been to replace single long test formats by multiple shorter test formats which assess a selection of the competences covered by the long test format.[24-26] The crucial factor for the success of this strategy would be to ensure that the shorter tests do indeed cover complete competences, i.e. the validity of the assessment must be safeguarded.

Research question 3: What is the reliability and validity of global performance ratings in an undergraduate clerkship with increased rater-student interactions? (chapter 4)

A clerkship was designed which differed from traditional clerkships in increased supervision, but also in reduced duration (3 weeks). Eight supervisors used global performance ratings (GPRs) on a five point Likert scale to assess the competence of all students who completed the clerkship. The expected higher reliability and validity of the GPRs due to increased supervision were not found. However, reliability and validity were not lower than reported for other assessment formats performed over a short testing time.[9,27-31] These results underscore that it is not enough to observe students a few times to arrive at a fair judgement, a finding that runs counter to the belief of many supervisors that a few observations suffice as a basis for a fair judgement. Nevertheless, because GPRs assess different domains of competence compared to more objective assessment formats, like OSCEs, they can be an additional source of information on aspects of students' clinical competence.[32-34] Given that GPRs alone are unlikely to achieve sufficient reliability and validity in undergraduate clinical education, it is advisable to combine GPRs with reliable assessment formats for specific competences. The combination of GPRs with the in-training assessment programme presented in this thesis may lead to an integrated clerkship assessment programme for feasible, reliable and more valid assessment of a broad range of competences.

Methodological considerations

We asked future examiners to comment on the implementation of the assessment formats studied in chapters 2, 3 and 4. Perhaps greater involvement in the implementation process of the future examiners, specifically members of staff and residents with experience in assessment, might have been advisable. In-depth interviews with members of staff and residents prior to constructing the assessment formats might have yielded useful information to

guide the construction of the assessment formats. Furthermore, repeating the in depth interviews when the assessment formats were being implemented might have resulted in early detection of problems and thus early intervention to resolve them.

In constructing the in-training assessment programme we might have considered other combinations of assessment formats (chapter 3). Taking account of specific aspects of the in-training assessment programme could have led to other combinations of assessment formats. Firstly, we might have considered the influence of the relative weights of each of the components of the in-training assessment programme on the reliability calculations. This might have led to improved reliability by balancing the weights of the different assessment formats, e.g. by implementing only multiple-sample formats. Secondly, combining GPRs with the in-training assessment programme could have broadened the range of competences assessed and could have improved reliability further. When the in-training assessment programme was introduced, however, the GPRs in the clerkship were changed to enable inclusion of in-training assessment results in the GPRs in an attempt to improve the reliability of the GPRs and thus the GPRs were no longer a separate assessment format. Therefore we could not combine the GPRs as a separate assessment format in the in-training assessment programme, which in retrospect was unfortunate.

The design of the GPR study contained aspects that counteracted each other (chapter 4). Intensified supervision was introduced to make supervisors better acquainted with the students and thus enhance the reliability of GPRs. Unfortunately, the potentially beneficial effect of intensified supervision was counteracted by the short duration of the three-week clerkship. However, feasibility considerations precluded intensified supervision in a clerkship of longer duration. A way to improve the study design with respect to reliability might have been to conduct the study in the clerkship in which the in-training assessment programme was introduced and ask examiners to give a GPR at the end of this clerkship (ten weeks).

Examiners did not receive any special training in assessing long case clinical examinations or the formats in the in-training assessment programme. Training might have had some advantages. Firstly, even though the literature offers no hard evidence to this effect, it is generally assumed that training of examiners can improve inter-rater agreement and thus the reliability of an assessment format.[35-38] Secondly, training might have had a beneficial effect on the feasibility of the new assessment formats, because training sessions offer the opportunity for a thorough introduction, explanation and discussion of the new programme, which might have enhanced examiners commitment to the assessment programme.

Conclusions for feasibility and reliability aspects of in-training assessment

Measures to improve reliability did not adversely affect the feasibility of the multiple long case clinical examinations or the in-training assessment programme. Feasibility remained sufficient to enable broad use of the performance based assessment formats in undergraduate clerkships and did not seem to be the main problem (chapters 2 and 3). This is not surprising, seeing that in constructing the assessment methods we had paid special attention to feasibility aspects. Prior to the implementation of the assessment programme, future examiners were asked to give their opinion on the feasibility of certain test formats. These opinions were incorporated into the design of the assessment methods. The educational programme of the clerkship was also scrutinised and test formats were carefully scheduled to fit into the educational programme. In summary, feasibility of new performance based assessment formats seems to benefit when account is taken of future examiners' opinions regarding feasibility and care is taken that assessment does not interfere with the educational programme. However, feasibility might improve even further with intensified examiner involvement and by piloting the assessment formats prior to the definitive implementation.

Reliability of the multiple long cases and of the in-training assessment programme was reasonably good due to the introduction of multiple assessment moments, multiple examiners and rating forms that brought more structure to the assessment.[9,14,39,40] The reliability of the GPRs remained low despite specific measures to improve it. Future research on in-training assessment in clerkships should target reliability. First of all research should address the use of a mix of test formats combining assessment of specific competences with more global assessments, such as GPRs.[15,34,41] Secondly, reliability might be improved by increasing the number of assessments within test formats, e.g. replacing single long assessments by multiple shorter ones.[9,24-26] When multiple shorter assessments are introduced, attention should be paid not only to feasibility and reliability aspects, but also to the validity for competence assessment, i.e. coverage of complete competences. An example of an assessment method that uses short assessments to assess complete competences is the mini-CEX in postgraduate education. The examiner observes the candidate and immediately afterwards completes a short rating scale instrument on a set of generic competence indicators. The mini-CEX can evaluate a wide range of learners' competences in a wide variety of clinical settings and with a diverse set of patient problems. Multiple encounters with different examiners and patients have been found to produce reliable data.[34,42]

In the studies described in chapters 2, 3 and 4 competences in which professional and personal skills play an important part were neither completely (long cases and in-training assessment programme) nor explicitly (GPRs) assessed. Combining in-training assessment with GPRs may improve the assessment of such competences, although large sample sizes will be needed to provide stable estimates of these competences.[43-45] Moreover, although GPRs may contribute to the assessment of these competences, it will be difficult to discriminate between specific competences. GPRs usually reflect a general impression of competence, including social and personal aspects, but they do not provide specific judgements on specific competences.[19,31,32]

Recently, portfolio assessment was introduced as a method to assess competences including professional and personal skills.[46,47] Combining the above mentioned assessment programmes with portfolio assessment is also a strategy that merits further exploration. E.g., the results of the tests could be collected in a portfolio which should be discussed on several occasions in the course of a clerkship.

Recommendations for feasibility and reliability aspects of in-training assessment

Careful consideration of both feasibility and reliability aspects appears to be helpful in constructing performance-based assessment formats and is therefore recommended. In addition, it should be recommended to critically appraise any measures that might optimise the assessment formats. These measures might target the assessment formats as well as aspects of implementation, such as examiner training in the use of test formats.

A mix of test formats, both multiple test formats and observed formats, is recommended to achieve reliable and feasible assessment of a broad range of competences. It would also be worthwhile to further explore the impact on reliability, feasibility and validity of assessment programmes that combine a mix of test formats and in which time consuming single test formats are replaced by multiple shorter sample assessments. Finally, including global performance ratings and portfolio assessments into these assessment programmes offers interesting possibilities for enhancing reliability and broadening the range of competences that can be assessed.[46,47]

Effects of an in-training programme on the learning environment

Findings and discussion

Research question 4: What are the frequency of supervision, feedback and assessment regarding a set of specified competences and what are the differences between disciplines? (chapter 5)

In all three clerkships studied, supervision (particularly observation) was sparse for all competences, feedback was mainly provided by residents and assessment mostly targeted patient-related competences. For all variables, the variation between students exceeded that between disciplines. This is in line with earlier research on skills training.[1,3,4,48-50] The findings of these studies suggest that even though competence learning is increasingly being emphasised in undergraduate and graduate education, optimal competence learning in the workplace remains yet to be achieved.

In the nineties, to enhance teaching of procedural and technical skills, many medical schools set up skills laboratories where students could practise skills on models and on each other.[51,52] These skills laboratories provided supervision, feedback and assessment of students' skill performance. This meant that for skill training students were not completely dependent on the workplace, where supervision, feedback and assessment were sparse. For competence learning, however, skills laboratories can only be part of the solution. The core characteristic of competences is that they should be performed in an authentic context. They cannot be learned through laboratory training only. 'Competence laboratory' is therefore a contradiction in terms. Students can prepare for aspects of a competence in a laboratory, but in order to learn to perform the whole competence, they need to practise in an authentic context that can only be offered in the workplace. Therefore a good learning environment in the workplace where supervision, feedback and

assessment of competences is provided, is even more important for competence learning than it is for skills learning. For competence learning to be successful, improvement of the workplace as a learning environment is an absolute necessity.

Research question 5: What is the effect of the in-training assessment programme on the frequency of supervision and the quality of feedback on clinical competences and on inter-student differences in these respects? (chapter 6)

From the viewpoint that assessment is a potentially powerful drive for students' and possibly also teachers' educational behaviour, we expected the introduction of an in-training assessment programme to affect the frequency of supervision and the quality of feedback as well as the differences between students in these aspects.[5,6,53] However, after the implementation of the in-training assessment programme no statistically significant increase in either variable was found. This suggests that the introduction of a new assessment programme for specific competences does not automatically result in an increase in supervision and feedback on the assessed competences nor in a decrease of the differences between students. The question naturally arises why this should be so. A possible explanation may be that the curriculum in action differs from the intended curriculum.[51] Elements of the programme in action of which we were unaware may have played a crucial role in bringing about the (non) effects of the assessment programme.[7,54] Another explanation may be that behavioural changes are most likely to occur when as many factors as possible are explicitly targeted.[55] The new assessment programme was implemented with the intention to change certain aspects in the clerkship. Additional measures aimed at supporting changes in these aspects, such as feedback training for supervisors, would most probably have facilitated behavioural changes.[56,57] If such training had been provided, the assessment could have triggered new behaviour, which could then have been facilitated by what supervisors had

learned in the training sessions. The conclusion seems justified that implementing an assessment programme and expecting it to change educational circumstances in a clerkship without monitoring the assessment programme or taking additional measures to support change offers a low chance of success.

> ***Research question 6: How is the in-training assessment programme actually carried out (curriculum in action) and what are the main features of supervision and feedback in the undergraduate clinical clerkship in which the in-training assessment programme is integrated? (chapter 7)***

In light of the findings of the study described in chapter six, the in-training assessment programme in action and its effects on supervision and feedback was explored qualitatively. The findings indicated that the assessment programme in action differed from the intended assessment programme.[51] Monitoring the implementation of new assessment programmes appears thus necessary, especially when the objective of the assessment is not only to assess student performance, but also to achieve certain educational goals.[7] The findings also revealed that both students and supervisors had difficulty obtaining and giving supervision and feedback, respectively. This qualitative research thus yielded outcomes that might be used to develop additional measures to support and facilitate changes in the learning environment. Such measures might include feedback training for examiners and scheduling dedicated time in the day-to-day practice schedule for providing feedback to students. In addition to training, designating time in the daily work routine for educational purposes will be necessary to achieve the full potential of changes that can be triggered by a new assessment programme. However, any additional measures to improve education will inevitably require additional time, effort and resources. In academic and teaching hospitals, patient care and research are often regarded as the two main tasks.[58] Education of undergraduate and graduate students comes third in the hierarchy of tasks and time constraints often

cause this educational task to fall victim to the competition with other duties.[59] This means that additional educational efforts will not be accepted easily. In the short term, educational efforts imply loss of time that might otherwise be devoted to patient care or research and only in the long term may educational efforts yield some profit in the form of well educated doctors and good clinical care. Furthermore, the hospital as a workplace is mainly equipped to provide patient care. Although patient care and education are not necessarily in conflict with each other, some effort and flexibility will be needed to create a workplace that fosters optimal performance in both patient care and teaching.[1,60] This suggests that creating an adequate learning environment in the workplace that is geared to the acquisition of competences will depend not only on educational measures, but also on strategic measures aimed at raising education to a status that is comparable with that of patient care and research.

Methodological considerations

The two quantitative studies presented in this section suffer from methodological drawbacks one way or the other.

The first shortcoming concerns the small sample sizes (chapters 5 and 6). Students rotate in clerkships in small groups. In some weeks only one student may enter the clerkship. Therefore, a long survey period is needed to reach acceptable sample sizes. However, long surveys were not always possible due to practical problems such as changes in educational programmes and rotating supervisors and examiners (residents and staff). Although Wolfhagen *et al*. reported that a sample size of ten student ratings was sufficient for reliable inferences from programme evaluation data, a small sample size is generally regarded as a methodological weakness.[61]

The second drawback is that the results of these studies were obtained by asking students to complete a questionnaire at the end of the clerkship rotation. The appropriateness of both the design (retrospective questionnaire) and the instrument (content of the

questionnaire) can be questioned. On one of the last days of the rotation, students were asked to complete a questionnaire which asked about the supervision and feedback they had received and the assessment of competences. The results may have been disproportionately influenced by experiences in the last weeks of the clerkship, whereas educational moments in the first few weeks of the clerkship may have become less vivid in students' memory or may even have been entirely forgotten. A different study design, in which students were asked to record the competences addressed in supervision, feedback and assessment in a log book over the course of the rotation could theoretically bypass these disadvantages. Information obtained in this way may be more accurate. On the other hand, the disadvantages of working with logs may also hamper adequate data gathering.[62] As for the content of the questionnaire, the items on feedback could have been more adequate. The questionnaire asked students to indicate which supervisor provided most information on a specific competence. However, this question elicited hardly any information on the true frequency of feedback. For example, if the groups of supervisors provided hardly any feedback, the one group that provided just a little bit more would be mentioned here. Conversely, the same holds for a situation in which all groups of supervisors provided abundant feedback.

Thirdly, the study design used in chapter 5 is rather weak. It is a pre-post design in which two different groups of students were studied in two different periods of time (prior to and after the implementation of the in-training assessment programme). This design makes it difficult to attribute any differences that were found between the two groups to the measure of interest, because other relevant factors may also have changed between the first and the second period. Moreover, the pre-intervention group was very small and standard deviations were high. This means that large differences in outcomes were needed to reach statistical significance. Changing the study design to a controlled experimental design and increasing the number of participants would increase the chance of detecting any effects of the in-

training assessment programme on the learning environment. However, it is not easy to use a controlled experimental design in an environment like that of a clerkship. The learning environment in a clerkship is very diverse and susceptible to influences from different groups of participants, such as patients, ancillary staff, residents and medical staff.[63] It is often the residents and the medical staff who have to carry out intended changes, i.e. use a different assessment format or supervise specific competences. It is very difficult to train supervisors and/or examiners to the extent that they all provide the same 'treatment'.[36] On top of that, the group of 'treatment' providers is constantly changing due to patient turnover, day and night shifts, residency rotations and even, though less frequently, staff rotations. There is also a very high variation among individual students in the education they receive in the clerkship. This means that, even if the 'treatment' would be provided in a consistent fashion, exposure to the treatment would vary among students in frequency.[64,65] The fact that per week at most one to two students enter a clerkship rotation means that a study would have to continue for several years to reach bigger sample sizes. Although the study design is weakened due to difficulties in providing the same treatment to all students, high individual variations between students and small sample sizes, these aspects are at the same time typical features of clerkship learning. Maybe the best approach to research into clerkship learning would be to perform small experiments to try and unravel the effects of measures implemented in the complex environment of a clerkship. Furthermore, considering the complexity of the setting, large randomised experiments of curricula are 'unlikely to ever be informative'.[66] Clerkships may be seen as a challenge inviting researchers to conquer the difficulties they present and unravel the secrets of workplace learning.

The qualitative study (chapter 7) revealed some serious problems concerning the in-training assessment programme. Although the results of this study are very useful, they can be regarded as coming 'a day after the fair' for both themes in this thesis. If the qualitative results had been obtained immediately after the implementation of the in-training assessment programme,

measures could have been taken to alter some aspects of the in-training assessment programme to improve both the programme and its effects on the learning environment. In that case the quantitative study on the effects of the in-training assessment programme on the learning environment might have yielded different results.

We investigated the curriculum in action (chapters 5, 6 and 7) and its effects on the learning environment. We did not investigate the learned curriculum, i.e. the effects of our interventions on students' competences (student outcome).[51] Although, the availability of feasible and reliable performance based assessment methods would make it feasible to obtain outcome measures for some competences at least, the complex learning environment and the high variation among individual students would preclude any unequivocal causal inferences to be made from student outcomes.

Conclusions regarding the effects of an in-training assessment programme on the learning environment

The low frequencies that were found for both supervision and feedback on competences in all three studies are striking and worrying, considering that workplace learning is of vital importance for acquiring competences and that supervision and feedback are vital components of an effective learning environment in the workplace. When it is mainly left to the students to judge what they still have to learn and to design their own programmes for competency learning, serious problems seem inevitable. Students are described as being incapable of making an adequate assessment of their own competences and also as showing a tendency to overestimate their own level of performance.[67-70] Therefore supervision by others besides students themselves is a prerequisite for an adequate judgement of students performance levels. Supervision should preferably be followed by clear and adequate feedback combined with opportunities for the student to practise the competence and demonstrate his or her progress at a later date.[70] Although we do not know exactly how much supervision and feedback students need to attain an adequate level

of competence, it is obvious that the low frequencies that were found in this thesis are glaringly insufficient.

In all three studies, patient related competences received most attention in supervision and feedback. It seems logical to emphasise these skills because they are highly content specific, which means that successful demonstration of direct patient-related competences in the context of one health problem offers little guarantee of success in demonstrating such competences in the context of the next problem.[14,23,71] Sustained attention for these competences is therefore necessary and ideally, this would mean that these competences are dealt with in relation to every health problem that can be encountered. However, competences involving communication skills also entail a certain degree of content specificity.[72,73] Furthermore, communication skills, professional skills and personal skills have been reported to decline over the course of medical education when they are not subject of regular supervision.[74-78] The fact that malpractice lawsuits are often triggered by complaints regarding communication skills, professional skills and personal skills suggests that continuous attention for competences in which these skills play a part is as important as attention for directly patient related competences.[79,80]

In the two quantitative studies we performed, senior supervisors were assumed to provide better quality feedback compared with junior supervisors, i.e. residents (chapters 5 and 6). Although there is no direct evidence to support this assumption, there is some circumstantial evidence. High quality feedback requires sufficient medical competence to provide reflection on performance and, for beginners, direction on performance. Furthermore, teaching skills such as setting learning objectives and providing structure and continuity in supervision and feedback are necessary to create a positive environment for feedback.[70] Residents are described as having deficiencies in their own clinical competences as well as in their teaching skills.[81-83] It seems therefore reasonable to entertain some doubts as to the quality of residents' feedback.[84]

Recommendations regarding the effects of an in-training assessment programme on the learning environment

When studying the effects of assessment on the learning environment in clerkships it is important to design a study and instruments that enable the close monitoring of these effects. Piloting designs and instruments may help prevent mistakes.

The learning environment for achieving competences in undergraduate clerkships is amenable to improvement. There is room for improvement regarding the frequencies of both supervision and feedback as well as the range of competences on which supervision and feedback are provided. Implementing new assessment programmes to improve the quality of clerkships as a learning environment is not enough. In the educational field careful monitoring of any newly introduced measures is of vital importance. Also any intended changes, such as more frequent feedback or follow-up on feedback, need to be firmly supported. Such support might be provided in the form of supervisor training and scheduling dedicated time for feedback. Monitoring, training and planning require effort and time and thus money. This means that, in addition to educational aspects, political and financial aspects need to be addressed if we are to achieve the change of culture and climate in which education is considered to be an important and rewarding task.

Final conclusions and implications

This thesis addresses the effects of an in-training assessment programme on the quality of the learning environment in relation to competence learning in an undergraduate clerkship as well as feasibility and reliability aspects of the in-training assessment programme.

The results demonstrate that it is possible to implement assessment methods that combine feasibility and reliability. Furthermore, many assessment formats can be combined to form a

mixed programme with adequate reliability and feasibility. Such a programme has the advantage of being able to reliably assess a far broader range of competences than each assessment format on its own. Such a broad assessment programme can yield professional judgements based on accumulated and triangulated information.[85] As for the effects of the in-training assessment programme on the learning environment, we found that clerkships are not the optimal learning environment for competence learning and the assessment programme did not have a detectable effect on the learning environment. Close monitoring of the assessment programme in action and supporting and facilitating intended change, e.g. by supervisor training and encouraging discussions on the position of education as one of the three tasks in hospitals are important measures that might improve the effects of assessment on the learning environment in clerkships.

The fact that assessment drives learning is more or less common knowledge in the field of medical education. This thesis shows that even when assessment in clerkships is improved, the assessment programme in action and its effects on the learning environment cannot be taken for granted. Attention should be given to all aspects of the learning environment in which the assessment is implemented and to identification of measures that might support the intended effects. Recommendations for the future are that attention should be focused on further improvement of assessment programmes but also on ways to enhance the learning environment in such a way that assessment programmes can achieve more educational goals than 'merely' improved assessment of student competence.

Implementing educational changes in clerkships is a difficult task. What may be considered an ideal solution from an educational viewpoint may be hard to achieve given the demands made on doctors and students in the day-to-day routine of the hospital as a workplace. Although in this thesis much emphasis was placed on the feasibility of measures, it turned out that this was not enough to successfully bridge the gap between the educational ideal and the practically achievable. Doctors in the workplace often feel that they

are forced to adopt new educational behaviours not because they want to do so, but merely because some educationalists have come up with another new idea. The intrinsic drive of doctors to change educational practice in the workplace is often lacking. How should we deal with this? First and foremost by explaining clearly and repeatedly why 'something new' is needed. It is important to show that change is not introduced simply because somebody came up with a new educational approach, but because there is evidence that education in the workplace is far from optimal. In the second place, the stakeholders in medical education should start to get to know each other. This means that educationalists should take an interest in what happens in the workplace, in how people work, what they think of education and which are the educational possibilities and impossibilities in a specific workplace. For doctors and students it is important to know how the learning environment can be improved and which factors are important in relation to supervision, feedback and assessment. In the third place, all parties involved should seriously consider each others motivations and make an inventory of potential improvements both on the part of educationalists and on the part of doctors and students. If educationalists are able to really support doctors and students in resolving the problems they encounter, this might engender faith in the educational effort and heighten motivation for active engagement in the implementation of new measures. Common interests can also be used as starting points for effective work.

In conclusion, the successful implementation of educational change is predicated on well informed, explicitly involved and well motivated participation of educationalists, doctors and students in the introduction of new educational formats and in the evaluation of the effects of these innovations. Only when working together real change can happen.

References

1. Irby DM. Teaching and learning in the ambulatory care setting, a thematic review of the literature. *Acad Med* 1995;70:898-931.
2. Regan-Smith M, Young WW, Keller AM. An efficient and effective teaching model for ambulatory education. *Acad Med* 2002;77:593-9.
3. Remmen R, Denekens J, Scherpbier A, Hermann I, van der Vleuten CPM, van Royen P, Bossaert L. An evaluation study on the didactic quality of clerkships. *Med Educ* 2000;34:460-4.
4. Van der Hem-Stokroos HH, Scherpbier AJJA, van der Vleuten CPM, de Vries H, Haarman HJThM. How effective is a clerkship as a learning environment? *Med Teach* 2001;23:608-13.
5. Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. *Med Educ* 1983;17:165-71.
6. Stillman PL, Haley HL, Regan MB, Philbin MM. Positive effects of a clinical performance assessment program. *Acad Med* 1991;66:481-3.
7. Van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ Theory Pract* 1996;1:41-67.
8. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990;65:S63-7.
9. Swanson DB. A measurement framework for performance-based tests. In: Hart IR, Harden RM, eds. *Further Developments in Assessing Clinical Competence.* Montreal: Can Heal Publications 1987;13-45.
10. Van der Vleuten CPM, Scherpbier AJJA, Dolmans DHJM, Schuwirth LWT, Verwijnen GM, Wolfhagen HAP. Clerkship assessment assessed. *Med Teach* 2000;22:592-600.
11. Turnbull J, MacFadyen J, van Barneveld C, Norman G. Clinical work sampling: a new approach to the problem of in training evaluation. *J Gen Int Med* 2000;15:556-61.
12. Feletti G, Cameron D, Dawson-Saunders B, des Groseilliers JP, Dooley B, Farmer E, McAvoy P. In-training assessment. In: Newble D, Jolly B, Wakeford R, eds. *The Certification and Recertification of Doctors: Issues in the Assessment of Clinical Competence.* Cambridge: Cambridge University Press 1994;151-66.
13. Turnbull J, van Barneveld C. Assessment of clinical performance: in-training evaluation. In: Norman GR, van der Vleuten CPM, Newble DI, eds. *International Handbook of Research in Medical Education.* Dordrecht/ Boston/London: Kluwer Academic Publishers 2002;793-810.
14. Van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardised patients: state of the art. *Teach Learn Med* 1990;2:58-76.

15. Van Luijk SJ, van der Vleuten CPM. A comparison of checklists and rating scales in performance-based testing. In: Hart IR, Harden RM, Des Marchais J, eds. *Current Developments in Assessing Clinical Competence.* Montreal: Can Health Publications 1992;357-82.
16. Cunnington JPW, Neville AJ, Norman GR. The risk of thoroughness: reliability and validity of global ratings and checklists in an OSCE. *Adv Health Sci Educ Theory Pract* 1997;1:227-33.
17. Van der Vleuten CPM. Validity of final examinations in undergraduate medical training. *BMJ* 2000;321:1217-9.
18. Wass V, Jolly B. Does observation add to the validity of the long case? *Med Educ* 2001;35:729-34.
19. Hull AL, Hodder S, Berger B, Ginsberg D, Lindheim N, Quan J, Kleinhenz ME. Validity of three clinical performance assessments of internal medicine clerks. *Acad Med* 1995;70:517-22.
20. Schuwirth LW, Southgate L, Page GG, Paget NS, Lescop JM, Lew SR, Wade WB, Baron-Maldonado M. When enough is enough: a conceptual basis for fair and defensible practice performance assessment. *Med Educ* 2002; 36:925-30.
21. Wass V, McGibbon D, van der Vleuten CPM. Composite undergraduate clinical examinations: how should the components be combined to maximize reliability? *Med Educ* 2001;35:326-30.
22. Wass V, van der Vleuten CPM. The long case. *Med Educ* 2004;38:1176-80.
23. Elstein AS, Shulman LS, Sprafka SA. *Medical Problem Solving: an Analysis of Clinical Reasoning.* Cambridge, MA.: Harvard University Press 1978.
24. Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. *Ann Int Med* 1995; 123:795-9.
25. Shatzer J, Wardrop J, Williams R, Hatch T. The generalizability of performance on different station length standardized patient cases. *Teach Learn Med* 1994;6:54-8.
26. Shatzer J, DaRosa D, Colliver JA, Barkmeier L. Station-length requirements for reliable performance-based examination scores. *Acad Med* 1993; 68:224-9.
27. Dielman TE, Hull A, Davis WK. Psychometric properties of clinical performance ratings. *Eval Health Prof* 1980;3:103-17.
28. Maxim BR, Dielman TE. Dimensionality, internal consistency and inter-rater reliability of clinical performance ratings. *Med Educ* 1987;21:130-7.
29. Metheny WP. Limitations of physician ratings in the assessment of student clinical performance in an obstetrics and gynecology clerkship. *Obstet Gynecol* 1991;78:136-41.
30. Callahan CA, Erdmann JB, Hojat M, Veloski JJ, Rattner S, Nasca TJ, Gonnella JS. Validity of faculty ratings of students' clinical competence in core clerkships in relation to scores on licensing examinations and supervisors' ratings in residency. *Acad Med* 2000;75:S71-3.

31. Norcini JJ, Webster GD, Grosso LJ, Blank LL, Benson JAJr. Ratings of residents' clinical competence and performance on certification examination. *J Med Educ* 1987;62:457-62.
32. Streiner DL. Global rating scales. In: Neufield VR, Norman GR, eds. *Assessing Clinical Competence*. New York: Springer Publishing Company 1985;114-41.
33. Keynan A, Friedman M, Benbassat J. Reliability of global rating scales in the assessment of clinical competence of medical students. *Med Educ* 1987,21:477-81.
34. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teach Learn Med* 2003;15:270-92.
35. Newble DI, Hoare J, Sheldrake PK. The selection and training of examiners for clinical examinations. *Med Educ* 1980;14:345-9.
36. Bennett KJ, Sackett DL, Haynes RB, Neufeld VR, Tugwell P, Roberts R. A controlled trial of teaching critical appraisal of the clinical literature to medical students. *JAMA* 1987;257:2451-4.
37. Van der Vleuten CPM, van Luijk SJ, van Ballegooijen AMJ, Swanson DB. Training and experience of examiners. *Med Educ* 1989;23:290-6.
38. Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clincal competence: a randomized trial. *Ann Intern Med* 2004;140:874-81.
39. Anastakis DJ, Cohen R, Reznick RK. The structured oral examination as a method for assessing surgical residents. *Am J Surg* 1991;162:67-70.
40. Van Ham I, Gerritsma J. The assessment of clinical competence in general practice with chart stimulated recall. In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP, eds. *Teaching and Assessing Clinical Competence*. Groningen: BoekWerk Publications 1990;306-9.
41. Van der Vleuten CPM, Norman GR, de Graaff E. Pittfalls in the pursuit of objectivity: issues of reliability. *Med Educ* 1991;25:110-8.
42. Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: a method for assessing clinical skills. *Ann Int Med* 2003;138:476-81.
43. Carline JD, Paauw DS, Thiede KW, Ramsey PG. Factors affecting the reliability of ratings of students' clinical skills in a medicine clerkship. *J Gen Int Med* 1992;7:506-10.
44. Carline JD, Wenrich MD, Ramsey PG. Characteristics of ratings of physician competence by professional associates. *Eval Health Prof* 1989;12:409-23.
45. Conway JM, Huffcutt AI. Psychometric properties of multisource performance ratings: a meta-analysis of subordinate, supervisor, peer and self-ratings. *Human Performance* 1997;10:331-60.
46. Lonka K, Slotte V, Halttunen M, Kurki T, Tiitinen A, Vaara L, Paavonen J. Portfolio as a learning tool in obstetrics and gynaecology undergraduate training. *Med Educ* 2001;35:1097-8.

47. MacLeod RD, Parkin C, Pullon S, Robertson G. Early clinical exposure to people who are dying: learning to care at the end of life. *Med Educ* 2003;37:51-8.
48. Jolly BC, Macdonald MM. Education for practice: the role of practical experience in undergraduate and general clinical training. *Med Educ* 1989;23:189-95.
49. Remmen R, Denekens J, Scherpbier AJJA, van der Vleuten CPM, Hermann I, van Puymbroeck H, Bossaert L. Evaluation of skills training during clerkships using student focus groups. *Med Teach* 1998;20:428-31.
50. Kassebaum DG, Eaglen RH. Shortcomings in the evaluation of students' clinical skills and behavior in medical school. *Acad Med* 1999;74:842-9.
51. Remmen R. *An Evaluation of Clinical Skills Training at the Medical School of the University of Antwerp* [thesis]. Antwerp: University of Antwerp 1999.
52. Hao J, Estrada J, Tropez-Sims S. The clinical skills laboratory: a cost-effective venue for teaching clinical skills to third-year medical students. *Acad Med* 2002;77:152.
53. Messick S. The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* 1994;23:13-23.
54. Van Luijk SJ, van der Vleuten CPM, Schelven RM. The relation between content and psychometric characteristics in performance-based testing. In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP, eds. *Teaching and Assessing Clinical Competence.* Groningen: BoekWerk Publications 1990;202-7.
55. Mann KV. Continuing medical education. In: Norman GR, van der Vleuten CPM, Newble DI, eds. *International Handbook of Research in Medical Education.* Dordrecht/Boston/London: Kluwer Academic Publishers 2002;793-810.
56. Dennick R. Long-term retention of teaching skills after attending the Teaching Improvement Project: a longitudinal, self-evaluation study. *Med Teach* 2003;25:314-8.
57. Morrison EH, Hafler JP. Yesterday a learner, today a teacher too: residents as teachers in 2000. *Pediatrics* 2000;105:238-41.
58. Stark P. Teaching and learning in the clinical setting: a qualitative study of the perceptions of students and teachers. *Med Educ* 2003;37:975-82.
59. Lemp H, Seale C. The hidden curriculum in undergraduate medical education: qualitative study of medical students' perception of teaching. *BMJ* 2004;329:770-3.
60. Regan-Smith M, Young WW, Keller AM. An efficient and effective teaching model for ambulatory education. *Acad Med* 2002;77:593-9.
61. Wolfhagen, HAP. *Kwaliteit van Klinisch Onderwijs* [thesis]. Maastricht: Maastricht University 1993.
62. Dolmans D, Schmidt A, van der Beek J, Beintema M, Gerver WJ. Does a student log provide a means to better structure clinical education? *Med Educ* 1999;33:89-94.

63. Woolliscroft JO. Medical student clinical education. In: Norman GR, van der Vleuten CPM, Newble DI, eds. *International Handbook of Research in Medical Education.* Dordrecht/Boston/London: Kluwer Academic Publishers 2002;365-80.
64. Gruppen LD, Wisdom K, Anderson DS, Woolliscroft JO. Assessing the consistency and educational benefits of students' clinical experiences during an ambulatory care internal medicine rotation. *Acad Med* 1993;68:674-80.
65. Seabrook MA, Woodfield SJ, Papagrigoriadis S, Rennie JA, Atheron A, Lawson M. Consistency of teaching in parallel surgical firms: an audit of student experiences at one medical school. *Med Educ* 2000:34:292-8.
66. Norman G. RCT = results confounded and trivial: the perils of grand educational experiments. *Med Educ* 2003;37:582-4.
67. Mattheos N, Nattestad A, Falk-Nilsson E, Attstrom R. The interactive examination: assessing students' self-assessment ability. *Med Educ* 2004;38:378-89.
68. Edwards RK, Kellner KR, Sistrom CL, Magyari EJ. Medical students self-assessment of performance on an obstetrics and gynecology clerkship. *Am J Obstet Gynecol* 2003;188:1078-82.
69. Tousignant M, DesMarchais JE. Accuracy of student self-assessment ability compared to their own performance in a problem-based learning medical program: a correlation study. *Adv Health Sci Educ Theory Pract* 2002; 7:19-27.
70. Kilminster S, Jolly B, van der Vleuten CPM. A framework for effective training for supervisors. *Med Teach* 2002;24:385-9.
71. Norman GR, Tugwell P, Feigtner JW, Muzzin LJ, Jacoby LL. Knowledge and clinical problem solving. *Med Educ* 1985;17:344-56.
72. Razavi D, Delvaux N, Marchal S, de Cock M, Farvacques C, Slachmuylder J. Testing health care professionals' communications skills: the usefulness of highly emotional standardized role-playing sessions with simulators. *Psycho-oncology* 2000;9:293-302.
73. Van Thiel J, Kraan HF, van der Vleuten CPM. Reliability and feasibility of measuring interviewing skills using the revised Maastricht History Taking and Advice Checklist. *Med Educ* 1991;25:224-9.
74. Pfeiffer C, Madray H, Ardolino A, Willms J. The rise and fall of students' skill in obtaining a medical history. *Med Educ* 1998;32:283-8.
75. Engler CM, Saltzman GA, Walker ML, Wolf FM. Medical student acquisition and retention of communication interviewing skills. *Med Educ* 1981;56:572-9.
76. Coulehan J, Williams PC. Vanquishing virtue: the impact of medical education. *Acad Med* 2001;76:598-605.
77. Woloschuk W, Harasym PH, Temple W. Attitude change during medical school: a cohort study. *Med Educ* 2004;38:522-4.

78. Masson N, Lester H. The attitudes of medical students towards homeless people: does medical school make a difference? *Med Educ* 2003;37:869-72.
79. Schwab S, Streuli R. Extrajudicial expert assessment - what is the basis for complaints against internists? *Schweiz Rundsch Med Prax* 1999;88:1981-93.
80. Nebel EJ. Malpractice: love thy patient. *Clinical Orthop* 2003;407:19-24.
81. Fox RA, Ingham Clark CL, Scotland AD, Dacre JE. A study of pre-registration house officers' clinical skills. *Med Educ* 2000;34:1007-12.
82. Sachdeva AK, Loiacono LA, Amiel GE, Blair PG, Friedman M, Roslyn JJ. Variability in the clinical skills of residents entering training programs in surgery. *Surgery* 1995;118:300-8; discussion 308-9.
83. Busari JO, Scherpbier AJJA, van der Vleuten CPM, Essed GGM. The perceptions of attending doctors of the role of residents as teachers of undergraduate clinical students. *Med Educ* 2003;37:241-7.
84. Stern DT, Williams BC, Gill A, Gruppen LD, Woolliscroft JO, Grum CM. Is there a relationship between attending physicians' and residents' teaching skills and students' examination scores? *Acad Med* 2000;75:1144-6.
85. Schuwirth L, van der Vleuten C. Merging views on assessment. *Med Educ* 2004;38:1208-11.

SUMMARY

The subject of this thesis is in-training assessment in a clerkship rotation in undergraduate medical education. In-training assessment is a specific form of performance based assessment which consists of multiple structured and (if possible) observed assessments, documented immediately following assessment and conducted at different moments over the course of a clerkship.

This thesis focuses on the feasibility and reliability of in-training assessment formats in clerkship rotations and on the effects of a new in-training assessment programme on the learning environment in clerkship rotations.

Chapter 1 presents a description of the learning environment of clerkships and a historical overview of performance based assessment in undergraduate clerkships. Students spend a major portion of their medical training working and learning in clerkship rotations, which make a crucial contribution to competence learning. However, the learning environment in clerkships is complex and far from optimal. Supervision and feedback are infrequent and deficient in structure and continuity. Because assessment is known to exert an important influence on students' and teachers' educational behaviours, it can be employed to modulate the learning environment in clerkships. The main assessment method in clerkships is performance based assessment, because clerkships offer excellent opportunities to assess students' performance in authentic contexts. Unfortunately, the performance based assessment formats that have traditionally been used in clerkships suffered from low reliability and validity. Measures to improve reliability and/or validity, such as the use of structured assessments, multiple assessments and multiple assessors can improve reliability and validity but may at the same time jeopardise feasibility. This means that there is a need for performance based assessment formats which are both feasible and reliable.

This thesis explores whether a new in-training assessment programme can improve the learning environment in clerkships. Firstly, the feasibility and reliability of in-training assessment formats were studied. Subsequently, the effects of a new in-training assessment programme on the learning environment was explored.

We addressed the following specific research questions:
1. What is the feasibility and reliability of multiple long case clinical examinations in an undergraduate clinical clerkship? (chapter 2)
2. What is the feasibility and reliability of the in-training assessment programme for assessing students' clinical performance? (chapter 3)
3. What is the reliability and validity of global performance ratings in an undergraduate clerkship with increased rater-student interactions? (chapter 4)
4. What are the frequencies of supervision, feedback and assessment regarding a set of specified competences and what are the differences between disciplines? (chapter 5)
5. What is the effect of the in-training assessment programme on the frequency of supervision and the quality of feedback on clinical competences and on inter-student differences in these respects? (chapter 6)
6. How is the in-training assessment programme actually carried out (curriculum in action) and what are the main features of supervision and feedback in the undergraduate clinical clerkship in which the in-training assessment programme is integrated? (chapter 7)

Chapter 2 addresses the feasibility and reliability of multiple long case clinical examinations. A long case clinical examination in the outpatient clinic was scheduled in the last week of a clerkship rotation. A maximum of two examinations per day were scheduled for a maximum of five consecutive days. The assessment was each day assigned to a (single) different examiner. Student performance was rated in two ways. Firstly, a form was used to rate performance on a number of items, with the mean score being used as the performance score and secondly, a single global judgement was given. Feasibility was expressed as the frequency of long case examinations and the number of assessors per student. Reliability was calculated by means of a generalisability analysis using a person x long case design. The results demonstrated good feasibility of assessments of individual students on six to eight long cases by four to five examiners in one week. With this increased

number of long cases and assessors, reliability of the mean item scores was 0.62-0.69 and for the global judgements it was 0.72-0.78. It was concluded that a moderate increase in the number of long case clinical examinations and in the number of (different) assessors was feasible. Reliability increased (especially for the global judgements) compared to the single long case clinical examination. The substantial time-investment for the multiple long case clinical examinations was no different than that needed to achieve reliable assessment with other methods.

In **chapter 3**, a study is described which examined the feasibility and reliability of an in-training assessment programme. We introduced in a clerkship rotation an in-training assessment programme comprising five test formats: three single-sample formats (student-patient encounter, critical appraisal session and case presentation) and two multiple-sample formats (twelve case write-ups and four structured long cases). In all, the in-training assessment programme consisted of nineteen assessment moments. Student performance of the relevant competences for each test format was rated on a five point Likert scale. As a measure of the feasibility of the assessment programme, we calculated the percentage of students who managed to complete all nineteen assessments (the arbitrary benchmark was set at 66.6%). Generalisability analysis was used to estimate the reliabilities for all nineteen assessments and for the five different assessment formats, respectively. The findings indicated that 69% of the students who took part in the in-training assessment programme met the required number of nineteen assessments. Although feasibility was adequate, further improvement is possible and might be achieved by scheduling dedicated time for educational purposes for both examiners and students, especially for the four structured long case clinical examinations. The results of the two reliability estimations differed considerably. With nineteen different assessments as items, reliability was 0.81. This estimate was heavily influenced by the weight of the twelve case write-ups and the four long cases. Estimated on the basis of five test formats, reliability turned out to reach only 0.55, which was probably

attributable to low reliability of the three single-sample formats. Nevertheless, even this low estimate was superior to the reliability of conventional performance based assessment formats. Thus the in-training assessment programme appeared to be feasible and more reliable than conventional performance based assessment. Feasibility might be improved by scheduling time for education and reliability might increase when single sample formats are replaced by multiple sample formats.

In **chapter 4**, reliability and validity of global performance ratings (GPRs) are examined. Students' overall performance was rated on a five point Likert scale by supervisors (members of staff) in a clerkship with more intensive supervision than is generally provided in clerkship rotations. We found an inter-rater reliability of 0.41 (generalisability-coefficient), which is comparable to inter-rater agreements reported in the literature on undergraduate medical education. Twenty-five GPRs would be needed to reach sufficient reliability. The predictive validity of the GPRs for the in-training assessment programme was 0.32, which was slightly higher than the predictive validity reported in studies in both undergraduate medical education and residency training. Disattenuation estimates the correlation between a predictor and criterion measure that both have perfect reliabilities. Disattenuation can therefore provide information on the real correlation. Disattenuated predictive validity was considerably higher (0.59), which suggests that GPRs can make a positive contribution to the evaluation of students' performance. It was concluded that the reliability and validity of global performance ratings, even though they remained problematic, were no worse than those of other assessment formats that require the same amount of testing time. Because GPRs can be an important source of information about students' clinical competence, combining GPRs with other assessment methods, such as in-training assessment, may lead to improved integral clerkship assessment.

The frequency of supervision, feedback and assessment in relation to a set of competences was investigated in three different undergraduate clerkships as a measure of adequate workplace learning.

The differences between the clerkships were also examined. This study is described in **chapter 5**. Clerks in the departments of internal medicine, surgery and paediatrics completed a questionnaire asking them about supervision (for seven specific settings), feedback and assessment on a set of sixteen relevant competences. Supervision (especially observation) proved to be sparse for all competences. Most supervision focused on direct patient related competences. Residents provided most of the feedback. Assessment mostly targeted patient related competences and team working skills. For all variables, variation among individual students exceeded variation among disciplines. This limited the significance of inter-disciplinary differences, although the department of internal medicine was found to provide the most supervision and show the greatest staff involvement. The conclusions from this study were that the clerkships failed to offer adequate conditions for competence learning in the workplace and that there was a huge inter-student variation in learning experiences. It is evident that there is ample room for improvement in the frequency of supervision, feedback and assessment of competences as well as in the consistency of delivery of the educational programme to all students.

Chapter 6 presents a pre-post study into the effects of the in-training assessment programme on the frequency of supervision and the quality of feedback for a set of specified competences as well as on inter-student differences in supervision and feedback in an internal medicine clerkship. Students were surveyed on received supervision (for seven specific settings) and feedback in relation to a set of thirteen relevant competences. It was expected that an assessment programme that targeted a specific group of competences would increase the frequency of supervision and quality of feedback on those competences across clerks. However, the results demonstrated no significant increase on either variable after the implementation of the in-training assessment programme in the internal medicine clerkship. Despite some methodological weaknesses, this study suggested important recommendations for subjects of further research, such as the way the assessment

programme is actually conducted (assessment in action) and specific research methods that are better able to detect the effects of interventions in the workplace.

Chapter 7 describes a qualitative exploration of the in-training assessment programme in action and of the main features of supervision and feedback in an undergraduate clinical clerkship in which the in-training assessment programme was incorporated. Nine students and seventeen assessors (eight residents and nine members of staff) participated in individual semi-structured interviews. The findings revealed that the assessment programme in action differed from the intended assessment programme. This indicates that it is of vital importance to ensure close monitoring of the implementation of an in-training assessment programme in an undergraduate clinical rotation. The results with regard to supervision and feedback showed that assessors provided very little supervision and feedback and also failed to follow-up on feedback they had given. Although students said that they wished to receive more supervision and feedback, they hardly ever made explicit requests to supervisors. Both students and assessors focussed on a few specific competences. The range of competences included in the in-training assessment programme was integrated neither into student learning nor into the supervision and feedback provided by staff. When giving feedback, assessors rarely gave borderline or fail ratings. These findings suggest that the introduction into a clerkship of an in-training assessment programme that covers a broad range of competences does not automatically result in more attention being paid to these competences by students and supervisors. A prerequisite for measurable effects of assessment on the learning environment seems to be additional measures to facilitate the intended changes, such as training of supervisors.

In **chapter 8**, the findings of this thesis are discussed and recommendations for the future are made. Methodological considerations in relation to the studies presented in this thesis are presented as well as some alternative approaches. In order to enhance reliability and feasibility of the three performance based assessment formats studied, it is recommended that developers of

performance based assessments should take account of reliability and feasibility in constructing an assessment programme and that measures to optimise assessment formats should be critically considered before the formats are implemented. Further research should look into the possibilities of an assessment programme that consists of a mix of short multiple and observed test formats in combination with global performance ratings. The results of these tests could be collected in a portfolio which should be discussed on several occasions in the course of a clerkship. This approach might make it feasible to assess the complete range of relevant competences.

Careful design of both study and instruments is recommended for studies of the effects of new performance based assessment methods on the learning environment in clinical clerkships. When assessment is used to enhance the learning environment in clinical clerkships, it is recommended to ensure careful monitoring of the assessment programme in action and to take measures to support and facilitate the desired changes. Potentially helpful measures include the training of supervisors and dedicated time in supervisors' schedules for the provision of feedback.

In order to reach the goal of improved learning in clerkships, attention should not only be focused on educational aspects but also on political and financial aspects. Strategic measures must be taken to raise education to the same level of importance as patient care or research. Moreover, the gap between the educational ideal and what is realistically achievable in the workplace needs to be addressed. This creates a need for the provision of adequate information to all those involved (i.e., doctors, health care managers, educationalists) about the possibilities for educational improvements in medical practice. It is also important that educationalists and clinicians gain insight into each others fields of expertise and identify common problems in order to foster a common motivation and mutual support in improving the learning environment in clerkships.

SAMENVATTING

Dit proefschrift heeft als onderwerp ‘in-training assessment’ tijdens een co-assistentschap in de opleiding tot basisarts. ‘In-training assessment’ wordt gedefinieerd als meerdere gestructureerde en (waar mogelijk) geobserveerde toetsmomenten die onmiddellijk na het toetsmoment gedocumenteerd worden en die verspreid over het hele co-assistentschap plaatsvinden. In deze samenvatting wordt de Engelse term ‘in-training assessment’ gebruikt, daar er geen Nederlands equivalent voor deze term is. ‘In-training assessment’ is een specifieke vorm van ‘performance-based assessment’. ‘Performance based assessment’ omvat toetsvormen die het handelen van studenten in de praktijk toetsen. Wederom zal wegens het ontbreken van een Nederlands equivalent in deze samenvatting de term ‘performance based assessment’ worden gebruikt. Dit proefschrift behandelt de haalbaarheid en betrouwbaarheid van ‘in-training assessment’ in co-assistentschappen en het effect van een nieuw ‘in-training assessment’ programma op de leeromgeving in co-assistentschappen.

In **hoofdstuk 1** wordt de leeromgeving in co-assistentschappen beschreven en wordt een overzicht gegeven van mogelijke vormen van ‘performance based assessment’ die in co-assistentschappen toegepast kunnen worden. Een belangrijk deel van de basisartsopleiding wordt gevormd door co-assistentschappen. Hoewel het leren in deze co-assistentschappen een onmisbare schakel is in de verwerving van competenties, blijkt dat de leeromgeving in de co-assistentschappen complex is en verre van optimaal. In co-assistentschappen is de frequentie van supervisie en feedback laag. Ook is er vaak weinig structuur en continuïteit met betrekking tot supervisie en feedback. Toetsing heeft een belangrijke sturende invloed op het gedrag van studenten en docenten met betrekking tot het onderwijs. Hierdoor kan toetsing gebruikt worden om de leeromgeving in co-assistentschappen te beïnvloeden. ‘Performance based’ toetsing tijdens de co-assistentschappen is belangrijk omdat het toetsing betreft van het handelen van studenten in een authentieke context. De ‘performance based’ toetsvormen die traditioneel gebruikt werden in co-assistentschappen hebben echter een geringe betrouwbaarheid en validiteit. Maatregelen om de betrouwbaarheid

en/of de validiteit te verhogen, zoals het structureren van toetsing en het verhogen van het aantal toetsen en examinatoren, hebben weliswaar een gunstig effect op betrouwbaarheid en validiteit, maar kunnen de haalbaarheid van de toetsing in gevaar brengen. Er is dan ook behoefte aan 'performance based' toetsvormen die zowel haalbaar als betrouwbaar zijn.

Dit proefschrift exploreert welke effecten een nieuw 'in-training assessment' programma als onderdeel van een co-assistentschap heeft op de leeromgeving van dat co-assistentschap. Eerst zijn de betrouwbaarheid en de haalbaarheid van verschillende toetsvormen die geschikt zijn voor 'in-training assessment' bestudeerd. Vervolgens zijn de effecten van het nieuwe 'in-training assessment' programma op de leeromgeving onderzocht. De specifieke onderzoeksvragen van dit proefschrift luiden als volgt:
1. Wat is de haalbaarheid en de betrouwbaarheid van meerdere patiëntgebonden klinische examens? (hoofdstuk 2)
2. Wat is de haalbaarheid en de betrouwbaarheid van een 'in-training assessment' programma voor toetsing van het handelen van studenten (performance)? (hoofdstuk 3)
3. Wat is de betrouwbaarheid en de validiteit van globale beoordelingen van co-assistenten (global performance ratings) in een co-assistentschap met toegenomen beoordelaars-student interacties? (hoofdstuk 4)
4. Wat is de frequentie van supervisie, feedback en toetsing in relatie tot een groep specifieke competenties en wat zijn de verschillen tussen disciplines? (hoofdstuk 5)
5. Wat is het effect van het 'in-training assessment' programma op de frequentie van supervisie en de kwaliteit van feedback ten aanzien van een groep specifieke competenties en op de verschillen tussen studenten hierin? (hoofdstuk 6)
6. Hoe wordt het 'in-training assessment' programma feitelijk uitgevoerd (curriculum in actie) en wat zijn de kenmerken van supervisie en feedback in een klinisch co-assistentschap waarin het 'in-training assessment' programma is opgenomen? (hoofdstuk 7)

In **hoofdstuk 2** worden de haalbaarheid en de betrouwbaarheid van meerdere patiëntgebonden klinische examens beschreven.

Tijdens de laatste week van het co-assistentschap in de polikliniek werd op maximaal vijf opeenvolgende dagen tweemaal per dag een patiëntgebonden examen gepland. Iedere dag was er een andere examinator. Het handelen van de student (performance) werd op twee manieren beoordeeld. Ten eerste werd met behulp van een scoreformulier een beoordeling gegeven op een aantal specifieke items waarvan vervolgens het gemiddelde werd berekend en ten tweede werd een globaal oordeel gegeven. De haalbaarheid van deze aanpak werd onderzocht door het aantal patiëntgebonden examens en het aantal verschillende examinatoren per student te bepalen. De betrouwbaarheid werd berekend middels een generaliseerbaarheidsanalyse waarbij een persoon x patiëntgebonden examen design werd gebruikt. Uit de resultaten bleek dat afname van zes tot acht patiëntgebonden examens per week door vier tot vijf examinatoren goed realiseerbaar was. De betrouwbaarheid van de patiëntgebonden examens, berekend op basis van de gemiddelde itemscores, was 0.62-0.69 en de betrouwbaarheid, berekend op basis van de globale oordelen, was 0.72–0.78. Hieruit kan geconcludeerd worden dat een matige toename van zowel het aantal patiëntgebonden examens als het aantal verschillende examinatoren haalbaar was. In vergelijking met het enkelvoudige patiëntgebonden examen gaf de betrouwbaarheid een toename te zien, met name voor de globale oordelen. De benodigde tijdinvestering voor meerdere patiëntgebonden examens verschilde niet van de tijdinvestering die nodig is om met andere toetsmethoden tot betrouwbare resultaten te komen.

In **hoofdstuk 3** worden de haalbaarheid en de betrouwbaarheid van het ‘in-training assessment’ programma beschreven. Het ‘in-training assessment’ programma bestond uit vijf verschillende toetsvormen: drie toetsen die eenmalig werden afgenomen (student-patiëntcontact inclusief een volledige anamnese en lichamelijk onderzoek, referaat en patiëntenpresentatie) en twee toetsen die meerdere malen werden afgenomen (twaalf samenvattingen en vier patiëntgebonden examens). Het ‘in-training assessment’ programma maakte deel uit van een klinisch co-

assistentschap en bestond in totaal uit negentien toetsen. Bij iedere toetsvorm werden de relevante competenties op een vijfpunts Likertschaal beoordeeld. De haalbaarheid werd bepaald aan de hand van het percentage studenten dat alle negentien toetsen aflegde (de grens werd arbitrair op 66.6% gesteld). De betrouwbaarheid van het toetsprogramma werd bepaald met behulp van een generaliseerbaarheidsanalyse met eerst alle negentien toetsen en vervolgens de vijf verschillende toetsvormen als items. Uit de resultaten bleek dat 69% van de studenten die het 'in-training assessment' programma doorliepen, alle negentien toetsen hadden gedaan. Hoewel de haalbaarheid voldeed aan onze eisen, zou het inroosteren van onderwijstijd voor examinator en student, vooral voor de vier gestructureerde patiëntgebonden examens, de haalbaarheid kunnen vergroten. Als de betrouwbaarheid berekend werd voor de negentien toets-vormen was deze hoger dan wanneer de betrouwbaarheid berekend werd voor de vijf verschillende toetsvormen. De betrouwbaarheid op basis van de negentien toetsen was 0.81. Deze betrouwbaarheid werd grotendeels bepaald door het gewicht van de twaalf samenvattingen en de vier patiëntgebonden examens. Als bij de betrouwbaarheidsbepaling evenveel gewicht werd toegekend aan de vijf verschillende toetsvormen, was de betrouwbaarheid 0.55, hetgeen nogal teleurstellend is. Waarschijnlijk is deze uitkomst bepaald door het relatief grote gewicht dat bij deze berekening wordt toegekend aan de drie enkelvoudige toetsvormen. Niettemin was de betrouwbaarheid op basis van de vijf toetsvormen nog altijd beter dan de betrouwbaarheid van conventionele 'performance based' toetsen. De conclusie was dan ook dat het 'in-training assessment' programma haalbaar was en betrouwbaarder dan conventionele 'performance based' toetsen. De haalbaarheid kan verder verbeterd worden door het roosteren van onderwijstijd voor examinatoren en studenten en de betrouwbaarheid kan verder toenemen als de eenmalig afgenomen toetsvormen vervangen worden door toetsvormen die meerdere keren worden afgenomen.

In **hoofdstuk 4** wordt een onderzoek gepresenteerd naar de betrouwbaarheid en validiteit van globale beoordelingen (global

performance ratings) in een co-assistentschap met meer interactie tussen beoordelaar en student dan in reguliere co-assistentschappen. Acht supervisoren, allen stafleden, gaven een globale beoordeling op een vijfpunts Likertschaal voor elke co-assistent die het extra gesuperviseerde co-assistentschap had doorlopen. De interbeoordelaarsbetrouwbaarheid (generaliseerbaarheidcoëfficiënt) van de globale beoordelingen was 0.41. Dit is vergelijkbaar met wat in de literatuur gerapporteerd wordt over interbeoordelaars-betrouwbaarheid in klinische co-assistentschappen. Uit extrapolaties bleek dat voldoende betrouwbaarheid verkregen kon worden bij vijfentwintig globale beoordelingen. De predictieve validiteit van de globale beoordelingen voor het 'in-training assessment' programma was 0.32. Dit is enigszins hoger dan de predictieve validiteit van globale beoordelingen voor andere toetsvormen in co-assistentschappen en vervolgopleidingen. Disattenuatie schat de correlatie tussen voorspellende en referentie variabele als de meetfout in beide nul is. Disattenuatie kan dus extra informatie geven over de 'echte' correlatie. De gedisattenueerde predictieve validiteit was aanzienlijk hoger, namelijk 0.59. Deze bevinding suggereert dat globale beoordelingen een positieve bijdrage kunnen hebben aan de beoordeling van 'performance'. De conclusie was dat de betrouwbaarheid en validiteit van globale beoordelingen weliswaar problematisch blijven, maar dat de bereikte betrouwbaarheid en validiteit niet lager waren dan die van andere toetsvormen die in een vergelijkbaar tijdsbestek kunnen worden afgenomen. Omdat globale beoordelingen een belangrijke bron van informatie kunnen zijn over de competentie van de student, kan de combinatie van globale beoordelingen met andere toetsmethoden, bijvoorbeeld 'in-training assessment', leiden tot een verbeterde integrale toetsing in de co-assistentschappen.

In **hoofdstuk 5** wordt een onderzoek beschreven naar de frequentie van supervisie, feedback en toetsing in relatie tot een groep specifieke competenties als maat voor adequaat leren in stages. Dit onderzoek werd uitgevoerd in drie verschillende co-assistentschappen waarbij ook gekeken werd naar de verschillen tussen disciplines. Co-assistenten bij Inwendige Geneeskunde,

Heelkunde en Kindergeneeskunde vulden een vragenlijst in over ontvangen supervisie (in zeven specifieke settings), feedback en toetsing in relatie tot een groep van zestien relevante competenties. Het bleek dat de co-assistenten met betrekking tot alle onderzochte competenties weinig gesuperviseerd en vooral weinig geobserveerd werden. Direct patiëntgerelateerde competenties werden het meest gesuperviseerd. Feedback werd het meest frequent gegeven door arts-assistenten. De toetsing was vooral gericht op patiëntgerelateerde competenties en op het werken in teamverband. Voor alle variabelen gold dat de verschillen tussen individuele studenten groter waren dan die tussen de disciplines, waardoor de significantie van interdisciplinaire verschillen beperkt was. Niettemin bleek dat de meeste supervisie werd gegeven bij Inwendige Geneeskunde, waar ook de betrokkenheid van de staf bij het geven van feedback het grootst was. Er werd geconcludeerd dat de voorwaarden voor adequate competentieverwerving slechts in beperkte mate aanwezig zijn op de werkplek en dat er grote individuele verschillen bestaan tussen studenten wat betreft hun ervaringen in stages. Duidelijk is dat de frequentie van supervisie, feedback en toetsing van competenties en de eenvormigheid van het aangeboden onderwijsprogramma in de drie bestudeerde co-assistentschappen verbetering behoeven.

In **hoofdstuk 6** wordt een onderzoek beschreven waarin voor een groep specifieke competenties de effecten gemeten werden van het 'in-training assessment' programma op de frequentie van supervisie, de kwaliteit van feedback en de verschillen tussen individuele studenten. Gegevens werden verzameld met behulp van een schriftelijke enquête waarin studenten gevraagd werd voor een groep van dertien relevante competenties aan te geven hoeveel supervisie (in zeven specifieke settings) en feedback zij ontvangen hadden. De verwachting was dat een toetsprogramma dat toegesneden was op een specifieke groep competenties, de frequentie van supervisie en de kwaliteit van feedback aangaande de betreffende competenties zou doen toenemen. De resultaten gaven echter voor geen enkele variabele een significante toename te zien na invoering van het 'in-training assessment' programma in het co-assistentschap Inwendige

Geneeskunde. Hoewel deze studie enkele methodologische zwakheden kent, zijn de resultaten van belang als richtsnoer voor vervolgonderzoek. Belangrijke aandachtspunten voor verder onderzoek zijn het toetsprogramma 'in actie' (in tegenstelling tot het 'beoogde' toetsprogramma) en meer geëigende onderzoeksmethoden voor het bepalen van de effecten van interventies in een setting als de stageplek.

Hoofdstuk 7 beschrijft een kwalitatieve studie over het 'in-training assessment' programma 'in actie' en de hoofdkenmerken van supervisie en feedback in een klinisch co-assistentschap waarin het 'in-training assessment' programma was opgenomen. Negen studenten en zeventien examinatoren (acht arts-assistenten en negen stafleden) verleenden hun medewerking aan individuele semi-gestructureerde interviews. Uit de resultaten bleek dat het toetsprogramma 'in actie' afweek van het 'beoogde' toetsprogramma. Dit suggereert dat de implementatie van een 'in-training assessment' programma veelvuldig dient te worden gecontroleerd. Verder bleek dat de examinatoren de studenten weinig superviseerden. Ook gaven de examinatoren weinig feedback en werd er in vervolgcontacten tussen student en examinator nauwelijks aandacht besteed aan de effecten van eerder gegeven feedback. Hoewel studenten zeiden dat ze graag meer supervisie en feedback wilden, vroegen ze daar zelden om. Zowel studenten als examinatoren richtten zich op enkele specifieke competenties. De competenties die door het 'in-training assessment' programma werden getoetst, bleken niet geïntegreerd te worden in het leren van studenten en het geven van supervisie en feedback door examinatoren. Het predikaat twijfelachtig of onvoldoende werd zelden toegekend als de examinatoren feedback gaven. Uit deze bevindingen blijkt dat de introductie van een 'in-training assessment' programma voor een aantal competenties niet automatisch leidt tot meer aandacht voor deze competenties zowel van studenten bij het leren als van examinatoren bij het geven van supervisie en feedback. Voor meetbare effecten van toetsing op de leeromgeving lijken extra maatregelen nodig om de gewenste veranderingen in de leeromgeving in een co-assistentschap te

bevorderen. Daarbij kan bijvoorbeeld gedacht worden aan trainingen van supervisoren.

In **hoofdstuk 8** worden de bevindingen van dit proefschrift besproken en aanbevelingen gedaan voor de toekomst. Ook worden de methodologische keuzes en mogelijke alternatieven besproken. Met betrekking tot de betrouwbaarheid en haalbaarheid van de drie bestudeerde 'in-training assessment' toetsvormen in co-assistentschappen wordt aanbevolen om reeds in de ontwerpfase aandacht te besteden aan specifieke aspecten van haalbaarheid en betrouwbaarheid. Ten aanzien van de implementatie van toetsvormen verdient het aanbeveling om maatregelen die de toetsvormen kunnen verbeteren, kritisch te beschouwen en waar mogelijk mee te nemen. Er wordt voorgesteld om meer onderzoek te doen naar toetsprogramma's met een veelvoud van korte toetsen (waar mogelijk geobserveerd) die meerdere malen worden afgenomen, gecombineerd met globale beoordelingen. De resultaten van de toetsen kunnen worden verzameld in een portfolio, dat enkele malen gedurende het co-assistentschap dient te worden besproken. Door bovenstaande combinatie van toetsvormen kan het volledige scala aan relevante competenties worden getoetst.

Wat betreft de effecten van het 'in-training assessment' programma op de leeromgeving in de co-assistentschappen is het aan te bevelen veel zorg te besteden aan het ontwerp van het onderzoek en de instrumenten om de effecten op de leeromgeving te meten. Het invoeren van een nieuw toetsprogramma leidt niet automatisch tot de gewenste effecten op de leeromgeving. Wanneer toetsing wordt ingezet als middel om de leeromgeving te verbeteren is het belangrijk om inzicht te hebben in het toetsprogramma 'in actie'. Daarnaast is het aan te bevelen om maatregelen te nemen om gewenste effecten te bevorderen en te ondersteunen. Daarbij kan bijvoorbeeld gedacht worden aan training van supervisoren en het expliciet inplannen van tijd voor de supervisor om feedback te geven.

Om de leeromgeving in de co-assistentschappen te kunnen verbeteren zullen naast onderwijskundige aspecten ook politieke en financiële aspecten aandacht behoeven. Strategische maatregelen met als doel onderwijs net zo belangrijk te maken als patiëntenzorg

of onderzoek zijn noodzakelijk. Daarnaast heeft de kloof tussen het onderwijskundige ideaal en het praktisch mogelijke op de werkplek aandacht nodig. Daarvoor is een goede informatievoorziening over mogelijke onderwijskundige verbeteringen in de medische praktijk nodig. Verder is aandacht van alle betrokkenen (o.a. dokters, gezondheidszorg managers, onderwijskundigen) voor elkaars vakgebied en het identificeren van gemeenschappelijke problemen noodzakelijk om te komen tot een gezamenlijke motivatie en gezamenlijke inzet voor verbetering van de leeromgeving.

DANKWOORD

Op de voorkant van dit proefschrift staat mijn naam. Ik heb dit proefschrift inderdaad zelf geschreven maar ik heb het niet alleen geschreven. Een groot aantal mensen hebben me daarbij geholpen en gesteund. Gaarne neem ik de gelegenheid om deze mensen te bedanken.

Coen en Ab, begeleiders in Amsterdam. Ik noem jullie gezamenlijk daar de begeleiding van dit proefschrift duidelijk een gezamenlijke taak was. Ab, door jouw inzet kreeg ik de mogelijkheid onderzoek te doen binnen de afdeling Inwendige Geneeskunde van het VUmc. Je hebt er persoonlijk op toegezien dat ik onderzoek kon verrichten en gegevens kon verzamelen en je hebt je volle gewicht gebruikt om ook na je emeritaat, het verzamelen van gegevens door te laten gaan. Ik ben je daar dankbaar voor. Coen, jij hebt na het vertrek van Ab mijn begeleiding in het VUmc op je genomen. Je hebt me bij aanvang gewaarschuwd dat je 'niet gemakkelijk' zou zijn. Ik heb ervaren dat als je 'niet gemakkelijk' was je daar altijd een reden voor had en dat die reden een leereffect in zich droeg. Het was aan mij of ik iets met dat leereffect wilde doen of niet, daarin liet je me vrij. Onze gesprekken heb ik als leerzaam ervaren en ik betreurde het dat die gesprekken in het laatste jaar nauwelijks meer mogelijk waren daar je in Maastricht ging werken. Dank voor je begeleiding.

Cees en Albert, jullie vertegenwoordigen de Maastrichtse kant van de begeleiding. Cees, voor mij was je een vaderlijke begeleider. Tijdens de werkbezoeken bleek je warm en zorgzaam maar ook duidelijk en streng. Hoewel je af en toe genadeloos kritiek gaf, zorgde je er altijd voor dat er voortgang mogelijk bleef en alternatieven uitgewerkt konden worden. Zonder jouw steun had ik het eerst artikel niet kunnen schrijven. Door jouw creatieve aanpak heb ik het geleerd. Eerst meekijken met jou, dan samen schrijven, dan alleen maar onder strikte supervisie, vervolgens zelfstandig. Je heb het mij volgens de regelen der kunst geleerd. Ik ben je daar zeer dankbaar voor. Ik hoop je nog vaak te horen lachen en zal er alles aan doen daar een bijdrage aan te hebben. Albert, jouw begeleiding

concentreerde zich primair op de grote lijnen, pas als die stonden, verdiepte je je in de details. Je betrokkenheid bij mijn proefschrift verraste me regelmatig zoals wanneer ik plotseling een artikel uit de post trok met een briefje erop ‘is interessant voor je artikel, Albert’. Ik heb je bijdrage aan de begeleiding zeer gewaardeerd.

Rita, promotiemaatje, paranimf en veelzijdige collega. We hebben samen een hele weg afgelegd. Letterlijk in de vorm van onze reizen naar Maastricht en naar de congressen waar we werk presenteerden. Figuurlijk van onze eerste probeersels om ‘de promotielijn’ te beschrijven tot aan artikelen in Medical Education. De pieken en dalen van dit promotietraject waren eenzaam geweest als ik ze niet met jou had kunnen delen, bedankt!

Marc, ‘oude’ baas, onderzoeker van onderwijs en paranimf. Samen met Cylla heb je me aangenomen op het ALCO. Beiden hebben jullie me de gelegenheid gegeven om me te ontplooien op het gebied van onderwijs. Marc, ik ben blij dat je paranimf wilt zijn.

Marian, ik ben met dit proefschrift begonnen toen jij onderwijs-directeur was. Je hebt mijn initiatief van het begin af aan gesteund. Je hebt manieren bedacht om een beetje tijd vrij te maken voor onderzoek van onderwijs in een tijd dat er in het onderwijsinstituut geen onderzoek gedaan kon worden. Toen mijn baan zo druk werd dat ik ook in de avonden geen tijd meer had om aan het proefschrift te werken heb je me aangeraden een dag thuis te gaan werken. Ondanks hoge werkdruk binnen het instituut ben je nooit aan die dag gekomen. Jij hebt de voorwaarden geschapen zodat ik het kon doen. Jij bent voor mij de initiator van het onderzoek van onderwijs in het onderwijsinstituut. Dit proefschrift is daarom ook een beetje van jou.

Ronnie, als nieuwe baas en nieuwe onderwijsdirecteur heb jij me de gelegenheid gegeven het werk af te maken. Ik vind het een eer dat je zitting wilde nemen in de leescommissie.

Abel, jij bent als onderwijscoördinator van de afdeling Inwendige Geneeskunde en als lid en later voorzitter van de werkgroep Toetsing betrokken geweest bij alle onderzoeken die ik heb uitgevoerd. Jij hebt meegedacht over de opzet van enquêtes en toetsvormen. Tevens heb je je ingezet om de nieuwe manier van toetsing te integreren in de dagelijkse gang van zaken op de afdeling Inwendige Geneeskunde. Samen bedachten we hoe de gegevensverzameling optimaal te maken en te houden. Jij was degene die dan op de afdeling 'de boer opging'. Ik ben je erg dankbaar voor deze inzet.

Ron en Arno, jullie belichamen de statistische ondersteuning van dit proefschrift. Ron, jij dacht in alle berekeningen met me mee. Belde me op met kritische vragen en leverde altijd op tijd de benodigde gegevens aan. Arno, jij gaf Ron gelegenheid me te ondersteunen. Voorts nam je me met groot geduld bij de hand bij de uitleg van berekeningen en de interpretatie van gegevens. Beide maakten jullie statistiek van iets engs tot iets benaderbaars en overwinbaars. Dat jullie dat kunnen als statistici vind ik een bijzonder verdienste.

Mereke, met een bijzondere inzet heb jij het Engels in dit proefschrift onderhanden genomen. Jij las je in in de materie tot en met de statistiek toe. Als jij iets niet begreep bleek het niet goed geformuleerd. Je verbeterde dan ook niet alleen het Engels maar het hele stuk. Ik dank je voor je inzet om de tekst optimaal te maken.

Renée, jij hebt een grote bijdrage geleverd aan het kwalitatief onderzoek in dit proefschrift. Vooral het uitwerken van de interviews en het coderen was een grote klus die ik zonder jou niet makkelijk geklaard had. Het huis en de poezen waren welkome onderwerpen tussen het coderen door. Ik dank je voor je bijdrage en je gezelligheid.

Vele anderen zijn betrokken geweest bij het tot stand komen van dit proefschrift doordat zij hun hulp en steun hebben gegeven.

Ali, Nicolien en Rita jullie hebben zorg gedragen voor het verzamelen van enquêtes en beoordelingsformulieren bij respectievelijk Inwendige Geneeskunde en Heelkunde. Karin, jij voorzag me met de meest recent uitdraaien van de instroom van co-assistenten en hielp me de juiste studentgegevens bij elkaar te krijgen. Lenie en Annelies, jullie waren mijn ingang tot de medewerkers bij Inwendige Geneeskunde. Elly, jij hebt me geholpen in de eerste stappen met SPSS.
Thei en John, jullie hebben het onderzoek op de Kindergeneeskunde gesteund. Sven, jij hebt de laatste onderzoeken bij de Inwendige Geneeskunde gesteund.
Wim en Christoffel, jullie hielpen met de studentenquêtes en het inlezen daarvan. Patrick en Cor, jullie hielpen met het elektronisch beoordelingsformulier en scanden de foto's.
Marianne en Yvonne, jullie hebben met eindeloos geduld de gegevens van de A4-logboeken ingevoerd.
Merilee, jij hebt mijn allereerste artikel gecorrigeerd.
Susan en Theo, jullie hebben me ingewijd in de eerste beginselen van kwalitatief onderzoek. Henk, jij hebt je aan het begin van het promotieonderzoek ingezet om het onderzoek vlot te trekken.
Anette en Ferry, medemaatjes in het onderzoek van onderwijs, jullie waren betrokken en vroegen bij iedere gelegenheid naar mijn voortgang.
Dank jullie wel dat ik bij jullie terecht kon!

Dit proefschrift was niet tot stand gekomen zonder de bijdrage van stafleden, arts-assistenten en co-assistenten bij de afdeling Inwendige Geneeskunde. De groep stafleden die vanwege hun specifieke taken een actieve bijdrage heeft geleverd aan verschillende onderzoeken wil ik graag met name bedanken. Frans, Roos, Abel, Evelien, Prabath, Jan, Samyah, Yvo, Frank, Michiel, dank voor jullie inzet.
Voorts was het kwalitatieve onderzoek niet tot stand gekomen zonder de inzet van de arts-assistenten Marielle, Oanh, Azam, Rob, Stijn, Herman, Neelke en Annemarie. Het aantal studenten dat aan de totstandkoming van dit proefschrift heeft meegewerkt is te groot om individueel te bedanken. Aan allen mijn dank.

De leden van de leescommissie ben ik erkentelijk voor de tijd en aandacht die zij hebben besteed aan de kritische beoordeling van dit proefschrift.

Medewerkers en ex-medewerkers van het ALCO: Antoon, Wynanda, Arthur, Pieter, Lia, Meta, Els, Fere, Marianne en Marcel, zonder jullie steun, belangstelling, relativerende grappen en medeleven was het werken aan dit proefschrift een stuk lastiger geweest. Menno, jij hebt me in het bijzonder gesteund door gedurende de laatste jaren een deel van mijn taken over te nemen. Ook onze gesprekken over de zin van deze promotie maakten dat ik andere zaken in mijn leven niet uit het oog verloor. Marlies, jij bent bij het laatste stuk betrokken geweest en hebt me bijzonder goed geholpen met de eindsprint. Daarvoor mijn dank.

Carla, grote zus. Van ons tweeen ben jij degene die in de medische praktijk werkzaam is. Mijn eerste les in professioneel gedrag kwam van jou. Jij hebt me in het eerste jaar van mijn studie doen besluiten dat ik een arts en mens wilde zijn die mensen waardeert om wie ze zijn, niet om wat ze doen of om hun positie. Ik ben nog dagelijks blij met dat besluit.

Pap en mam, van jullie hebben we het doorzettingsvermogen dat ons beiden kenmerkt en dat nodig is om een proefschrift af te ronden. Als dochters van ‘ kleine zelfstandige ondernemers’ weten we wat werken is en zetten we ons optimaal in voor onze zaak. Pap, u hield niet van poespas dus ik hou het kort, ik hou van u. Mam, u steunde het initiatief om dit proefschrift te gaan schrijven. Die steun bleef ik voelen ondanks het feit dat ik minder tijd voor u had. Ik hoop de komende jaren veel momenten met u te delen.

Lieve Ud, dit proefschrift is nu klaar mede dankzij jouw steun en jouw bijdrage aan de vormgeving. Werd het op het eind toch nog een gezamenlijk project. Ik hou van je.

CURRICULUM VITAE

Hester Daelmans is geboren op 18 februari 1963 te Roermond. In 1981 voltooide zij haar VWO opleiding aan de Rijksscholengemeenschap te Roermond. In 1988 behaalde zij het artsexamen aan het Universitair Medisch Centrum St. Radboud te Nijmegen (destijds Rooms Katholieke Universiteit Nijmegen).
In 1989 begon zij te werken als assistent in opleiding bij de afdeling anatomie van de Vrije Universiteit te Amsterdam. Een onderdeel van dat werk was begeleiding van studenten tijdens de snijzaalpractica. Hester raakte geboeid door het onderwijs en vervolgde vanaf 1993 haar carrière bij het Opleidingscentrum van het VU-ziekenhuis (OCV) als vakdocent Interne Geneeskunde t.b.v. de verkorte A-opleiding en als vakdocent anatomie en ziekteleer t.b.v. de schakelopleiding voor allochtone studenten. In 1995 werd zij tevens werkzaam als ALCO-docent bij het onderwijsinstituut van de Vrije Universiteit. Daar professionaliseerde zij zich verder in onderwijs door een groot aantal cursussen te volgen van de Nederlandse Vereniging van Medisch Onderwijs (NVMO) (toen nog onder de vlag van Specialisatie Medisch Onderwijs). Het onderwerp toetsing interesseerde haar bijzonder. Zij schreef in 1997 een onderzoeksproject voor het Studeerbaarheidsfonds en verrichtte een onderzoek naar interbeoordelaarsbetrouwbaarheid van het anamnese-station in het stationsexamen op het ALCO. Zij publiceerde hierover in 1999 in het Tijdschrift Medisch Onderwijs (destijds Bulletin Medisch Onderwijs). Naast het ALCO-docentschap werd zij in 1997 werkzaam als secretaris van de werkgroep Toetsing die belast was met de herziening van de toetsing in de co-assistentschappen aan de Vrije Universiteit. Zij schreef uit hoofde van die functie in 1999 een projectaanvraag voor het Onderwijs Kwaliteits Fonds en verrichtte onderzoek naar nieuwe toetsvormen in de co-assistentschappen (opgenomen als hoofdstuk 3 in dit proefschrift). In 1998 werd zij hoofd ALCO en enkele jaren later voorzitter van de projectgroep klinische vaardigheden (PKV). In 1999 schreef zij het plan voor haar promotietraject en begon met het systematisch verrichten van onderzoek.
Sinds 2003 is zij naast haar functie als hoofd ALCO actief betrokken bij het ontwerp van het nieuw curriculum geneeskunde in het VUmc.

Haar aandachtsgebieden in het nieuw curriculum zijn: ontwerp en ontwikkeling van de competentielijn en toetsing van competenties. Voor het huidige curriculum geneeskunde heeft zij in 2004 als voorzitter PKV nieuw beleid aangaande patientgebonden vaardigheden geformuleerd. Daarnaast formuleert en bestendigt zij als voorzitter van de werkgroep ‘hepatitis B beleid opleiding geneeskunde’ en van de werkgroep ‘buitenlandse co-assistentschappen’ beleid op deze deelgebieden.